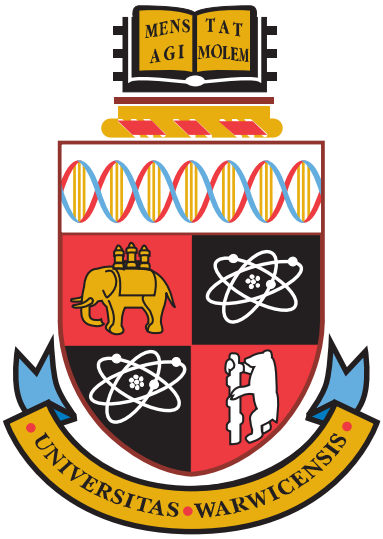


Unconstrained Face Recognition with Occlusions

by

Xingjie Wei

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Computer Science

August 2014

THE UNIVERSITY OF WARWICK

Contents

Declarations

I hereby declare that this dissertation entitled *Unconstrained face recognition with occlusions* is an original work and has not been submitted for a degree or diploma or other qualification at any other University.

Publications

Book Chapter

1. B. Arbab-Zavar, **X. Wei**, J.D. Bustard, M.S. Nixon and C.-T. Li, **On Forensic Use of Biometrics**, *Handbook of Digital Forensics of Multimedia Data and Devices*, ed. by A. T. S. Ho and S Li, John Wiley & Sons, Inc. 2014 (in press)

Journal

2. **X. Wei**, C.-T. Li, Z. Lei, D. Yi, Stan Z. Li, **Dynamic Image-to-Class Warping for Occluded Face Recognition**, *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2035-2050, December 2014

3. Y. Guan, **X. Wei**, C.-T. Li, **On the Generalization Power of Face and Gait in Gender Recognition**, *International Journal of Digital Crime and Forensics*, vol. 6, no. 1, pp. 1-8, January 2014

Conference

4. Y. Guan, **X. Wei**, C.-T. Li, and Y. Keller, **People Identification and Tracking through Fusion of Facial and Gait Features**, *Proceeding of International Workshop on Biometrics (BIOMET)*, 23-24 June 2014, Sofia, Bulgaria

5. X. Lin, **X. Wei** and C.-T. Li, **Two improved forensic methods of detecting contrast enhancement in digital images**, *Proceeding of IS&T/SPIE Conference on Media Watermarking, Security, and Forensics*, 2-6 February 2014, San Francisco, US

6. Y. Guan, **X. Wei**, C.-T. Li, G.L.Marcialis, F. Roli and M.Tistarelli, **Combining Gait and Face for Tackling the Elapsed Time Challenges**, *Proceeding of IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS'13)*, 29 September-2 October 2013, Washington DC, US

7. **X. Wei** and C.-T. Li, **Fixation and Saccade based Face Recognition from Single Image per Person with Various Occlusions and Expressions**, *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'13)*, 23-28 June 2013, Portland, Oregon, US

8. **X. Wei**, C.-T. Li and Y. Hu, **Face Recognition with Occlusion Using Dynamic Image-to-Class Warping (DICW)**, *Proceeding of IEEE International Conference on Automatic Face and Gesture Recognition (FG'13)*, 22-26 April 2013, Shanghai, China

9. **X. Wei**, C.-T. Li and Y. Hu, **Robust Face Recognition with Occlusions in both Reference and Query Images**, *Proceeding of International Workshop on Biometrics and Forensics (IWBF'13)*, 4-5 April 2013, Lisbon, Portugal

10. **X. Wei**, C.-T. Li and Y. Hu, **Robust Face Recognition under Varying Illumination and Occlusion considering Structured Sparsity**, *Proceeding of The International Conference on Digital Image Computing: Techniques and Applications (DICTA'12)*, 3-5 December 2012, Fremantle, Australia

Research Training

1. Transferable Skills courses: Teamworking and Networking, University of Warwick, Coventry, UK, April, 2011.

2. The 7th International Summer School on Pattern Recognition, Plymouth, UK, September, 2011.

3. The 9th Summer School for Advanced Studies on Biometrics: Understanding Man-Machine Interactions In Forensics and Security Applications (with grant), Alghero, Italy, June, 2012.

4. Communication & Impact for Female Early Career Researchers Course (with grant), Cumberland Lodge, Great Windsor Park and BBC Broadcasting House, UK, January, 2013.

5. The 10th Summer school for advanced studies on biometrics: How Biometrics meets Forensics, Security and the E-society Challenges of Tomorrow (Best presentation award), Alghero, Italy, June, 2013.

Abstract

Face recognition is one of the most active research topics in the interdisciplinary areas of biometrics, pattern recognition, computer vision and machine learning. Nowadays, there has been significant progress on automatic face recognition in controlled conditions. However, the performance in unconstrained conditions is still unsatisfactory. Face recognition systems in real-world environments often have to confront uncontrollable and unpredictable conditions such as large changes in illumination, pose, expression and occlusions, which introduce more intra-class variations and degrade the recognition performance. Compared with these factor related problems, the occlusion problem is relatively less studied in the research community.

The overall goal of this thesis is to design robust algorithms for face recognition with occlusions in unconstrained environments. In uncontrollable environments, the occlusion preprocessing and detection are generally very difficult. Compared with previous works, we focus on directly performing recognition with the presence of occlusions. We deal with the occlusion problem in two directions and propose three novel algorithms to handle the occlusions in face images while also considering other factors.

We propose a reconstruction based method *structured sparse representation based face recognition* when multiple gallery images are available for each subject. We point out that the non-zeros entries in the occlusion coefficient vector also have a cluster structure and propose a structured occlusion dictionary for better modelling them. On the other hand, we propose a local matching based method *Dynamic Image-to-Class Warping* (DICW) when the number of gallery images per subject is limited. DICW considers the inherent structure of the face and the experimental results confirm that the facial order is critical for

recognition. In addition, we further propose a novel method *fixations and saccades based classification* when only one single gallery image is available for each subject. It is an extension of DICW and can be also applied to deal with other problems in face recognition caused by local deformations.

The proposed algorithms are evaluated on standard face databases with various types occlusions and experimental results confirmed their effectiveness. We also consider several important and practical problems which are less noticed (i.e., coupled factors, occlusions in gallery or/and probe sets and the *single sample per person* problem) in face recognition and provide solutions to them.

Abbreviations

AAM	Active appearance model
ASM	Active shape model
AUC	Area under ROC curve
CCTV	Closed-circuit television
CMs	Coupled mappings
CR	Collaborative representation
DCT	Discrete cosine transform coefficients
DICW	Dynamic image-to-class warping
DMMA	Discriminative multi-manifold analysis
DTW	Dynamic time warping
EER	Equal error rate
EGM	Elastic graph matching
FAR	False acceptance rate
FBI	Federal bureau of investigation
FERET	Facial recognition technology evaluation
FRGC	Face recognition grand challenge
FRR	False rejection rate
FRVT	Face recognition vendor test
FSC	Fixations and saccades based classification
GMM	Gaussian mixture models

GP	Gabor phase
GSRC	Gabor-feature based SRC
HMM	Hidden Markov model
ICA	Independent component analysis
IGO	Image gradient orientation
MKD	Multi-keypoint descriptors
MRF	Markov random field
NBNN	Naive Bayes nearest neighbour
NIST	National institute of standards and technology
LARK	Locally adaptive regression kernel descriptor
LBP	Local binary pattern
LDA	Liner discriminant analysis
LFW	Labeled faces in the wild database
LHS	Local higher-order statistics
LLE	Locally linear embedding
LLN	Law of large numbers
LPQ	Local phase quantization
LRC	Linear regression classification
PCA	Principal component analysis
PD	Partial distance
PIE	Pose, illumination and expression
PIFS	Partitioned iterated function system
RLS	Regularised least square
ROC	Receiver operating characteristic
RRC	Regularised robust coding
RSC	Robust sparse coding
SOM	Self-organising map

SR	Super-resolution
SRC	Sparse representation based classification
SSPP	Single sample per person
SSR	Structured sparse reconstruction
SSS	Small sample size
SVDD	Support vector data description
SVM	Support vector machine
TFWM	The face we make database
WLD	Weber local descriptor

List of Tables

List of Figures

Chapter 1

Overview

1.1 Why face ?

With increasing emphasis on national and global security, there is a growing and urgent need for human identification. Biometrics is the science of identifying an individual based on the physiological and behavioural characteristics. It can be traced to 14th century China, where merchants used children's palm and footprints to distinguish them from one another. The physiological characteristics are related to the shape of the body including face, iris, retina, fingerprint, palmprint, palm vein, hand geometry, DNA, earlobe, etc. The behavioral characteristics are related to the pattern of behaviour of a person such as gait, signature, keystroke dynamics, voice, etc.

Among these biometric traits, face is the most commonly seen and used one in our daily life. Since the advent of photography, both government agencies and private organisations have kept face photo collections of people (e.g., personal identification documents, passports, membership cards). With the wide use of digital cameras, smart phones and CCTVs in the past decade, face images can be even more easily generated every day. In addition, nowadays these images can be rapidly transmitted and shared through the

highly developed social networks (e.g., Facebook[1], Flickr[2], Instagram[3]). The face is almost the most common and familiar biometric trait in daily life. Compared with other biometric traits such as fingerprint and iris, the face has several advantages as listed below that make it one of the most preferred biometric traits for human identification:

- *Biological nature*: The face is a very convenient biometric characteristic used by humans in the recognition of people, which makes it probably the most common biometric trait for authentication and authorization purposes. For example, in access control, it is easy for administrators to track and analyse the authorised person from his/her face data after authentication. The help from ordinary users (e.g., administrators in this case) can improve the reliability and applicability of the recognition systems, whereas fingerprint or iris recognition systems require an expert with professional skills to provide reliable confirmation.

- *Non-intrusion*: Different from fingerprint and iris collections, facial images can be easily acquired from a distance without physical contact. People feel more comfortable for using faces as identifier in their daily life. A face recognition system can collect biometric data in a user-friendly way, which is easily accepted by the public.

- *Less cooperation*: Compared with iris and fingerprint, face recognition has a lower requirement of subject cooperation. In some particular applications such as surveillance, a face recognition system can identify a person without active participation from the subjects.

1.2 Motivations

Face recognition is one of the most active research topics in the interdisciplinary areas of biometrics, pattern recognition, computer vision and machine learning. Nowadays, there has been significant progress on automatic face recognition in controlled conditions

[1] https://www.facebook.com/
[2] https://www.flickr.com/
[3] http://instagram.com/

[1]. However, the performance in unconstrained conditions is still unsatisfactory. Face recognition systems in real-world environments often have to confront uncontrollable and unpredictable conditions such as large changes in illumination, pose, expression and occlusions, which introduce more intra-class variations and degrade the recognition performance. Compared with the traditional PIE (i.e., pose, illumination and expression) problems, the occlusion related problem is relatively less studied in the research community. As analysed by the works in [2] and [3], occlusions can significantly decrease the performance of face recognition algorithms. We select the occlusion problem as the research topic since it is a common seen, less studied and extensible problem.

1.2.1 Face recognition with occlusions: commonly seen problem

In real-world environments, faces are easily occluded, which decreases the recognition accuracy. Generally speaking, faces can be occluded in passive and active ways. On the one hand, faces capture can be affected by environmental factors such as extreme illumination (responsible for strong shadows), limited field of view (causing partial faces), poor image quality (attributed to cause blurring) and objects in front of the face (e.g., food, mobile phone, others' faces). These factors are difficult to fully control in real-world environments. On the other hand, faces are usually occluded by the subjects themselves. For example, in the daily life, it is very common that people wear facial accessories such as sunglasses, scarves, hats, masks and veils for personal or cultural reasons (Figure 1.1a). In face recognition scenarios, one can ask the subject to get rid of these accessories when subject cooperation is applicable (e.g., border control). But this will give rise to inconvenience. What is more, in the security/crime related scenarios, people tend to use occlusions to hide their identities and subject cooperation is not applicable at all (e.g., surveillance). Figure 1.1b and Figure 1.1c are some example images captured in the *London Riots* in 2011 and the *Boston Marathon bombings* in 2013. Current commercial face recognition systems suffer a significant performance drop with these unconstrained, occluded face images [4]. Based on the analysis in [5], we summarise the categorisation of different sources of occlusions

in face images in Table 1.1. The research of the occluded face recognition problem is very important since it widely occurs in both daily life and security related scenarios.

Table 1.1: The categorisation of different sources of occlusions

Type	Scenario	Occlusion example
Passive	Extreme illumination	Strong shadows
	Limited field of view	Partial faces
	Poor image quality	Blurring, underexposure, overexposure
	External occlusion	Food, mobile phones, pets, others' faces
Active	Daily facial accessory	Sunglasses, scarves, hats, veils
	Self-occlusion	Non-frontal poses, hair, hands
	Criminal camouflage	Masks, sunglasses, caps
	Privacy protection	Anti-recognition make-up, cosmetics

1.2.2 Face recognition with occlusions: less studied problem

Face recognition with occlusions in unconstrained environments is relatively less studied in the research community yet. There are a large number or face recognition algorithms which are designed to deal with the PIE problems, however, relatively less work focuses on the occlusion problem. For the pose problem, the pose variations can be reduced by using face images with multi-view poses or 3D face models. For the illumination problem, a face image with a novel illumination can be represented with the combinations of face images with existing illuminations. For the expression problem, expression variations can be modelled with well-defined expression images (e.g., the six basic expressions: anger, disgust, fear, happiness, sadness and surprise). But different from that, a kind of occlusion (e.g., sunglasses) cannot be generated by other kinds of occlusions (e.g., scarf). The types of occlusions are unpredictable in practical scenarios and no prior knowledge is available. All of these make it difficult for the algorithms for dealing with other factors to be easily and directly applied to handle the occlusions. Research into the occluded face recognition problem is very important since it is an overlooked problem in face recognition.

(a)

(b)

(c)

Figure 1.1: Examples of occluded face images: occluded images in the daily life [a] (a), and occluded images in crime scenarios (e.g., London riots[b] (b), Boston Marathon bombings[c] (c)).

[a] Images from the Internet
[b] Images from the Metropolitan Police: http://content.met.police.uk/
[c] Images from the FBI http://www.fbi.gov/

1.2.3 Face recognition with occlusions: extensible problem

The research on the occlusion problem can help solve other emerging problems in face recognition. There are some emerging uncontrollable factors which will largely affect the face recognition performance such as heavy make-up [6, 7] and plastic surgery [8–10]. These problems share some similarities with the occlusion problem. For example, they are commonly seen in faces and some of them affect the face locally like occlusions. More importantly, similar to occlusions, these factors are difficult to predict. Collecting sufficient training samples is also not easy since these factors in testing images may be largely different from those in the training data. On the other hand, some local distortions by large expressions such as closed eyes and open mouth can also be seen as *occlusions* on the face. Due to these common characteristics, research into occluded face recognition problem can help to design algorithms for dealing with other factors in face recognition.

1.3 Objectives

Note that there are two related but different problems to face recognition with occlusions: *occluded face detection* and *occluded face recovery*. The first task is to determine whether a face image is occluded or not [11], which can be used for automatically rejecting the occluded images in applications such as passport image enrolment. This rejection mechanism is not always suitable for face recognition in some scenarios (e.g., surveillance) where no alternative image can be obtained due to the lack of subject cooperation. The second task is to restore the occluded regions in face images [12, 13]. It can recover the occluded areas but is unable to directly contribute to recognition since the identity information can be contaminated during inpainting.

An intuitive idea for handling occlusions in face recognition is to detect the occluded regions first and then perform recognition using only the non-occluded parts. Min *et al.* [14, 15] adopted the SVM classifier to detect the occluded regions in a face image and then used only the non-occluded areas of a probe face (i.e., query face) as well

as the corresponding areas of the gallery faces (i.e., reference faces) for recognition. But note that the occlusion types in the training images are the same as those in the testing images. Jia and Martínez [16, 17] used a skin colour based mask to remove the occluded areas for recognition. However, the types of occlusions are unpredictable in practical scenarios. The location, size and shape of occlusions are unknown, hence increasing the difficulty in segmenting the occluded regions from the face images. Currently most of the occlusion detectors are trained on faces with specific types of occlusions (i.e., the training is data-dependent) and hence generalise poorly to various types of occlusions in real-world environments.

The overall goal of this thesis is to design robust algorithms for face recognition with occlusions in unconstrained environments. There are several differences between our work and the above previous works in dealing with the occlusion problem:

1. We focus on performing recognition with the presence of occlusions. As mentioned before, occlusion detection is not always applicable in real-world environments. Recognising a face with occlusions directly is very practical in applications.

2. We consider the case that occlusions also exist in the gallery/training images. Most of the previous works assume that occlusions only occur in probe images. But in practical scenarios, the gallery images can also be contaminated by occlusions. In this thesis we do not make any assumption about the presence of occlusions (i.e., in gallery or probe) and consider that occlusions may occur in both gallery and probe images.

3. We deal with the occlusions while considering other factors (e.g., illumination, expression changes) which also affect the recognition accuracy. In the real-world scenarios, these factors are usually coupled. Most efforts of robust face recognition research are devoted to deal with each of the factors independently, but less work focuses on simultaneously handling them. In this thesis we do not only focus on the occlusion problem itself. We also consider other factors such as illumination/expression changes, rotations, small pose changes, etc.

1.4 Main contributions

We propose three novel algorithms in two directions (i.e., the reconstruction based and the local matching based, see Figure 1.2) to handle the occlusions in face images when performing recognition. In addition, we also take account of several practical problems (i.e., coupled factors, occlusions in gallery or/and probe sets, the *single sample per person* (SSPP) problem [18] and the *closed world* problem [19]) in face recognition and provide solutions to them. The main contributions are summarised in detail as follows.

1. We give an overview of the existing works for face recognition with occlusions. We carefully categorise these approaches and analyse their advantages and disadvantages, which provides a clear big picture of the state-of-the-art techniques. We summarise three occlusion cases which a face recognition system may encounter in the real-world applications but most of the current works do not consider.

2. We propose a reconstruction based method *structured sparse representation based face recognition* [20] when multiple gallery images are available for each subject. We point out that the non-zeros entries in the occlusion coefficient vector also have a cluster structure and propose a structured occlusion dictionary for better modelling them. In addition, we consider the coupled condition of extreme illuminations and occlusions, which is practical in real-world environments.

3. We propose a local matching based method *Dynamic Image-to-Class Warping* (DICW) [21–23] when the number of gallery images per subject is limited. Different from most of the existing works that simply treat occluded face recognition as a recovery problem or just employ the framework for general object classification, DICW considers the inherent structure of the face and our experimental results confirm that the facial order is critical for recognition. This method can be also applied to deal with other object recognition problems where the geometric relationship or contextual information of features should be considered.

4. We propose a novel method *fixations and saccades based classification* (FSC)

[24] for occluded face recognition when only one single gallery images is available for each subject (i.e., the SSPP problem). FSC is inspired by the observations in human visual perception. It is an extension of the aforementioned DICW and can be also applied to deal with other problems in face recognition caused by local deformations (e.g., the facial expression problem).

1.5 Outline

The rest of this thesis is organised as follows. Figure 1.2 shows the structure of the thesis.

Chapter 2 first discusses face recognition process performed by both human beings and computers. It then introduces the approaches of unconstrained face recognition and further reviews the state-of-the-art methods for occlusion problems and summarises the current challenges.

Chapter 3 first briefly introduces the sparse representation based classification (SRC) [25] model. Based on that, it explains the proposed approach *structured sparse representation based face recognition* in two steps: 1) the structured sparse representation (SSR) and 2) the structured occlusion dictionary. Then, it describes the strategy of the combination of SSR and the robust Weber local descriptor (WLD) [26] to handle coupled factors. At last, it provides the experimental analysis and the discussion for the parameters.

Chapter 4 introduces the proposed method *Dynamic Image-to-Class Warping* (DICW) step by step, from image representation, modelling to implementation. Three occlusion cases, namely: occlusions in the gallery set only, occlusions in the probe set only, and occlusions in both, are discussed. It confirms the effectiveness of DICW with extensive experimental results and also discusses its robustness to misalignment and other parameters of DICW.

Chapter 5 first introduces the SSPP problem and the background of human visual perception. It then explains the proposed method *fixations and saccades based classification* (FSC) in details. Finally it presents the experimental analysis as well as

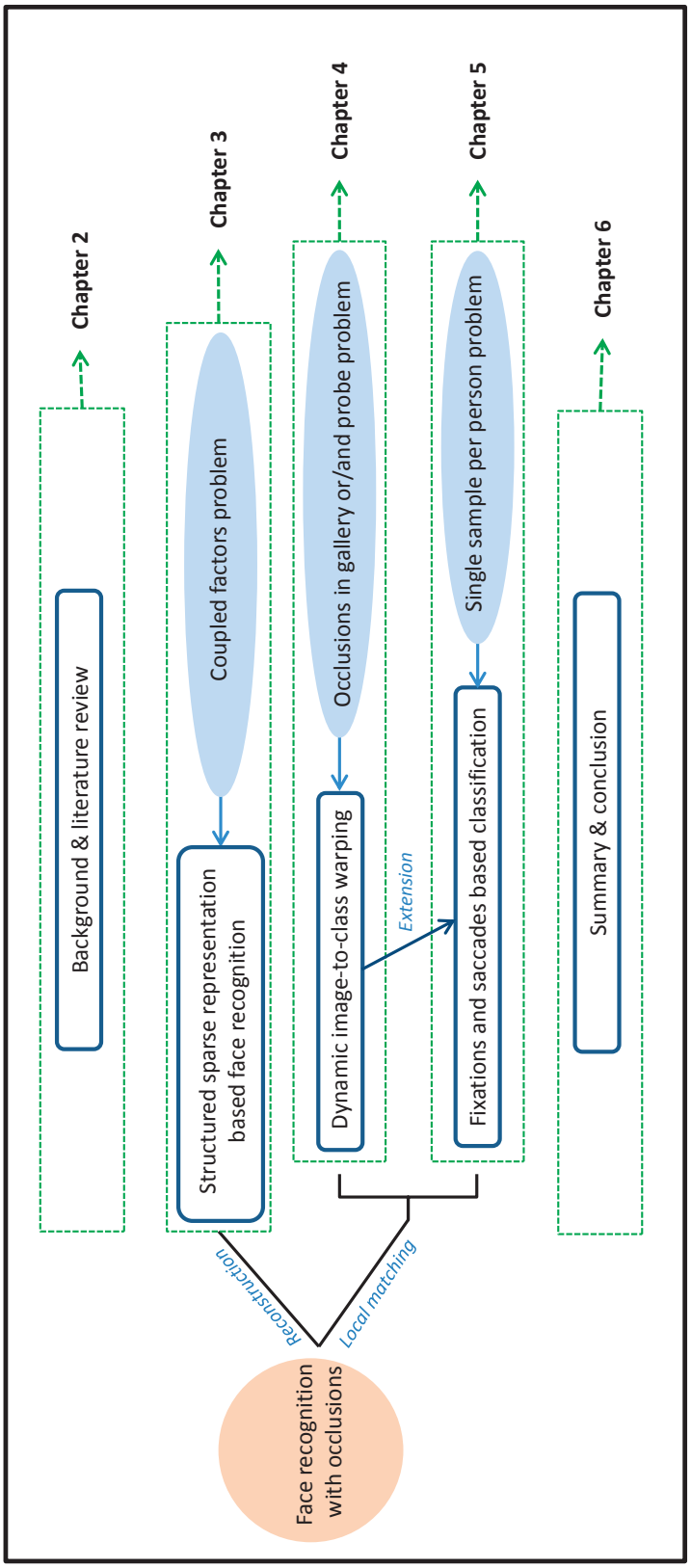


Figure 1.2: The structure of the thesis

discussions on the parameter.

Chapter 6 summarises the achievements of this thesis and presents some future research directions.

Chapter 2

Literature review

2.1 Face recognition by humans: forensic face recognition

In the past, before the advent of electronic computers, face recognition was already widely used in forensics. The first attempt we are aware of to identify a subject by comparing a pair of facial photographs was reported in a British court in 1871 [27]. In 1890, the French criminologist Alphonse Bertillon set forth a set of standards for forensic photography. He developed a taxonomy to describe some of the physiological features of the head and face, which is called *portrait parlé* or *spoken portrait* [28]. The combination of the anthropometric measurements and the *spoken portrait* developed by Bertillon is called *Bertillonage* and was fast adopted by the police and the judicial systems. Figure 2.1 demonstrates a principle of Bertillon's anthropometry.

In a typical forensic face recognition scenario, a forensic expert is given face images of a suspect (e.g., mug-shot images) and a questioned person (i.e., the perpetrator). The forensic expert will give a value which represents the degree to which these images appear to be of the same person. There are four main categories of face recognition approaches used by the forensic experts [30, 31]: holistic comparison, morphological analysis, anthropometry, and superimposition.

- *Holistic comparison.* In holistic comparison, faces are visually compared as a

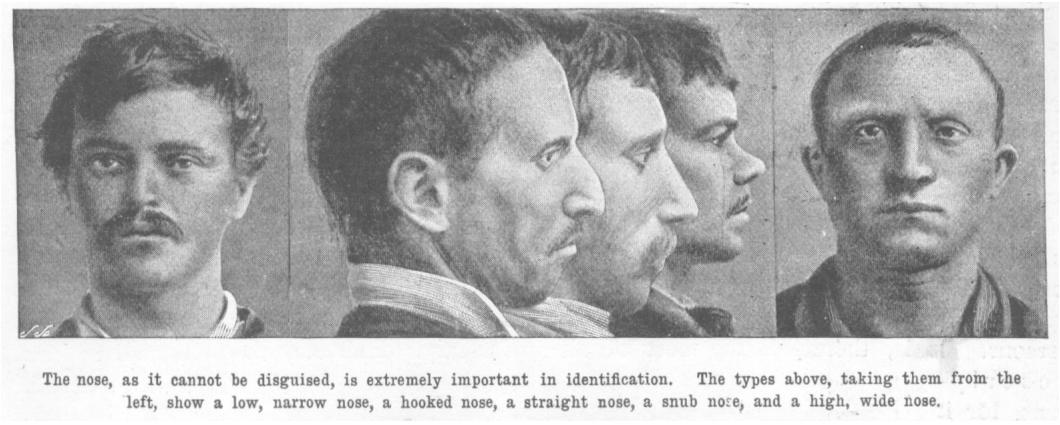


Figure 2.1: Face recognition by humans: an illustration of a principle of Bertillon's anthropometry. Image is excerpted from [29].

whole by the forensic experts. This is the simplest way and can be performed as a pre-step for other methods.

- *Morphological analysis*. In morphological analysis, the local features of the face will be analysed and compared by the forensic experts. They carry out an exhaustive analysis on the similarities and differences in observed faces, trait by trait on the nose, mouth, eyebrows, etc., even the soft traits such as marks, moles, wrinkles, etc. The location and distribution of local facial features are considered but not explicitly measured compared with anthropometry based approaches.

- *Anthropometry*. Anthropometry refers to the measurement of the human individual, which can be used for human recognition. Different from morphological analysis, in face anthropometry, the quantification measurements (e.g., spatial distance and angles) between specific facial landmarks (e.g., the tip of the nose, the centres of the eyes) are used for comparison.

- *Superimposition*. In superimposition, one face image is overlaid onto another and the forensic experts need to determine whether there is an alignment and correspondence of the facial features.

Aside from the area of forensics, face recognition has also received research interests from neuroscientists and psychologists, as well as computer scientists [32].

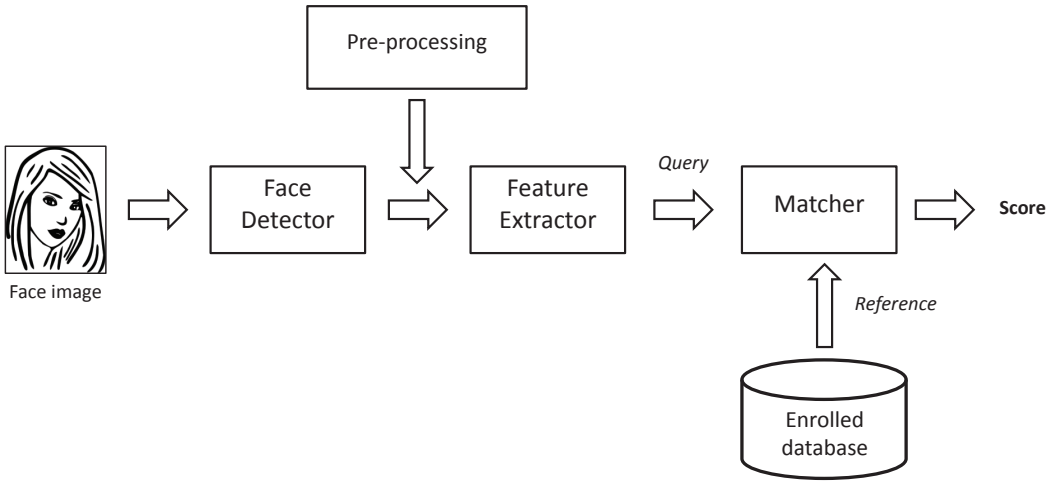


Figure 2.2: Framework of an automatic face recognition system.

Although the mechanism of face recognition by human visual system has not been fully understood yet, researchers are trying to analyse the observations and simulate the strategies in face recognition by humans for designing automatic face recognition algorithms. An human face recognition inspired algorithm is proposed in Chapter 5.

2.2 Face recognition by computers: automatic face recognition

A general automatic face recognition system usually consists of the following modules: a face detector, a feature extractor and a matcher (Figure 2.2). The face detector [33] crops the face area from the background of an image. The feature extractor then extracts effective information from face images for distinguishing different individuals. Usually pre-processing such as face alignment by the facial landmarks and face normalisation (e.g., scale, illumination condition) is performed before feature extraction. Then the matcher compares two faces (e.g., the probe and the gallery) based on the extracted features and then a similarity score is calculated. Face recognition is based on the similarity scores and its performance highly relies on the extracted features and classification algorithms used to distinguish faces.

2.2.1 Performance metrics

Face recognition consists of two main experimental modes: *identification* and *verification* [34]. In identification (i.e., one-to-many search), the face recognition system searches a database for an enrolled gallery sample to match the probe sample. An ordered list of the top *n* most similar matches are returned as the possible identities of the probe according to the similarity scores. The performance of the system in the identification mode is measured in terms of *rank-n identification rate* which is the rate at which the true association is included in the top *n* matches. The *identification rate* usually refers to rank-1 identification rate where the system returns a single match (the best match), as the most probable association with the probe sample. This rate is also called the *recognition rate*.

On the other hand, verification (i.e., one-to-one check) is the task where the recognition system attempts to confirm an individual's claimed identity by comparing the probe sample to the individual's previously enrolled sample. Verification is based on a decision threshold which is set by computing the similarity scores of all sample pairs in the gallery. The threshold is chosen to separate the genuine (i.e., match) similarity scores distribution from the impostor (i.e., non-match) similarity scores distribution and gives the best performance based on one of the following metrics [35]. Here an impostor is a person who submits a sample attempting to claim the identity of another person.

- *False acceptance rate* (FAR) is the rate at which the comparison between two different individuals' samples is erroneously accepted by the system as the true match. In other words, FAR is the percentage of the impostor scores which are higher than the decision threshold.

- *False rejection rate* (FRR) is the percentage of times when an individual is not matched to his/her own existing template. In other words, FRR is the percentage of the genuine scores which are lower than the decision threshold.

- *Equal error rate* (EER) is the rate at which both acceptance and rejection errors are equal (i.e., FAR=FRR). Generally, the lower the EER value, the higher the accuracy of the biometric system.

Both identification and verification tasks are used when evaluating general face recognition algorithms. For occluded face recognition, most of the works conduct experiments for the identification tasks since it is more commonly seen in the application scenarios. In the experiments of this thesis we mainly consider the identification task and use rank-1 rate for performance measurement. The verification is considered in Chapter 4 for face pair comparison tasks.

2.2.2 Brief history

In the early time, the main recognition approaches are geometric feature-based methods which rely on measurements between specific facial landmarks. This is similar to the anthropometry based methods mentioned in Section 2.1 in the face recognition by forensic experts. The first attempt to perform automatic face recognition started in 1965 by Chan and Bledsoe [36] in a semi-automated mode where a set of facial features were extracted from the photographs by humans. The first fully automatic face recognition system was presented by Kanade [37] in 1973, which marked a milestone at that time. Before 1990, it was the early stage of face recognition and most of the approaches were just *tested in the lab*.

In 1990s, appearance-based linear subspace analysis approaches and statistical models became the mainstream. Turk and Pentland [38] applied the Principal Component Analysis (PCA) on face images, which was referred to as *Eigenface*. These eigenfaces are the eigenvectors associated with the largest eigenvalues of the covariance matrix of the training samples, which ensures that the data variance is maintained while eliminating unnecessary existing correlations among the original features (i.e., dimensions). PCA based approaches greatly reduce the computational cost for high-dimensional data and inspire more active research in face recognition. Belhumeur *et al.* proposed *Fisherface* [39], which is based on the Liner Discriminant Analysis (LDA). This approach also performs dimensionality reduction while preserving as much of the class discriminatory information as possible. Other popular methods at that time include the Local Feature Analysis (LFA)

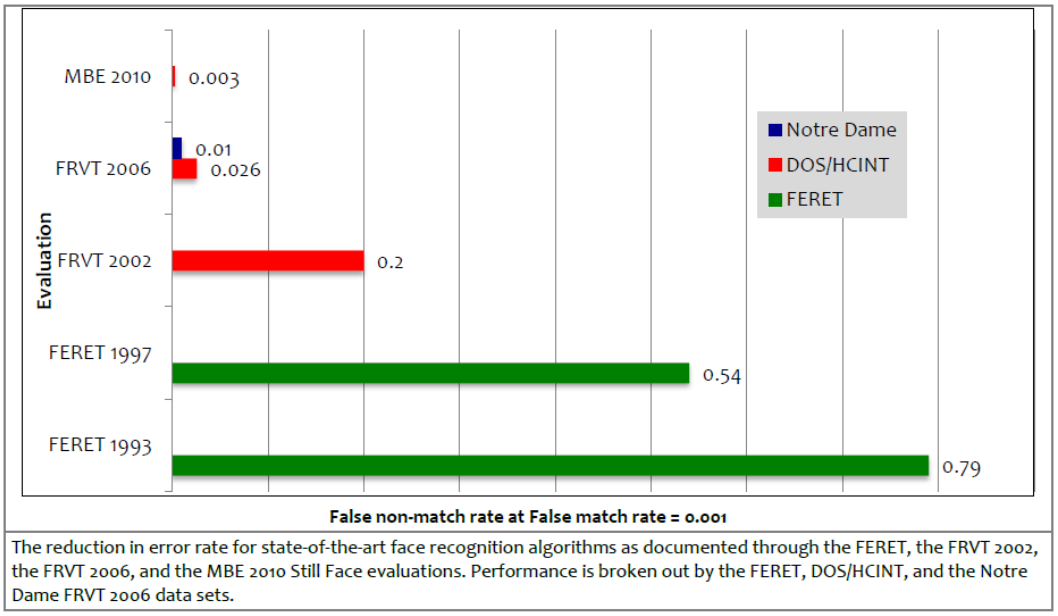


Figure 2.3: Progression of face recognition accuracy measurements. Image is excerpted from [45].

[40], the Elastic Graph Matching (EGM) [41], the Hidden Markov Model (HMM) [42], etc. Besides these, the statistical models such as Active Shape Model (ASM) [43] and Active Appearance Model (AAM) [44], which are widely used in face modelling and face alignment, also appeared in this period. Not only where numerous theoretical research works conducted but also commercial applications appeared in this period.

From the late 90s to present, research in face recognition has focused on uncontrollable scenarios (e.g., large pose, illumination and expression changes). Locally Linear Embedding (LLE) [46], illumination core model [47], 3D Morphable Model [48], Independent Component Analysis (ICA) [49], Local Binary Pattern (LBP) [50] and Sparse Representation based Classification (SRC) [25] are the representative methods in this period. The performance of automatic face recognition techniques have been evaluated in a series of large-scale tests conducted by the National Institute of Standards and Technology (NIST[1]). Good examples include, the Facial Recognition Technology evaluation (FERET) [51], the Face Recognition Vendor Test (FRVT) [1] and the Face Recognition Grand

[1]http://www.nist.gov/

Challenge (FRGC) [52]. Over the past decades, there has been significant progress in automatic face recognition. The false reject rate (FRR) of the best performing face recognition algorithm has decreased from 79% in 1993 to 0.3% in 2010 at a false accept rate (FAR) of 0.1% [53] (Figure 2.3).

In 2007, a benchmark database *Labeled Faces in the Wild*[2] for unconstrained face recognition was built. It contains more than 13,000 images (5,749 subjects) of faces collected from the Internet. The only constraint on these faces is that they were detected by the Viola-Jones face detector [55]. The face recognition accuracy by humans on this database is about 97.53% to 99.20% (Figure 2.4). And the accuracy for automatic face recognition algorithms has been improved from about 72.45% [56] in 2007 to more than 97.0% in 2014. Three algorithms/commerical softwares appeared in 2014 achieved the performance very close to or even better than humans (i.e., *Face++* [57]: 97.27%, *DeepFace* [58]: 97.35% and *GaussianFace* [59]: 98.52%). These results are very impressive. But currently it is too optimistic to say that the performance of machine is comparable or better than that of humans for face recognition. The images used in those tests are cropped faces excluding the external facial features such as hair and ear. However, humans are able to fully exploit external features and context information for recognition especially in difficult conditions (as shown in Figure 2.4).

During the *London Riots* in 2011, the London Metropolitan Police used automatic face recognition software in attempting to find the suspects, but in most cases it failed [60]. The poor quality of most CCTV footage makes it almost impossible to trust standard automatic facial recognition techniques. Changes in illumination, image quality, background and orientation can easily fool the face recognition systems (images are shown in Figure 1.1b). In 2013, Klontz and Jain [4] conducted a case study that used the photographs of the two suspects in the *Boston Marathon bombings* to match against a background set of mugshots. The suspects' photographs released by the FBI were captured in the uncontrollable environment and their faces were partially occluded by sunglasses and

[2]http://vis-www.cs.umass.edu/lfw/

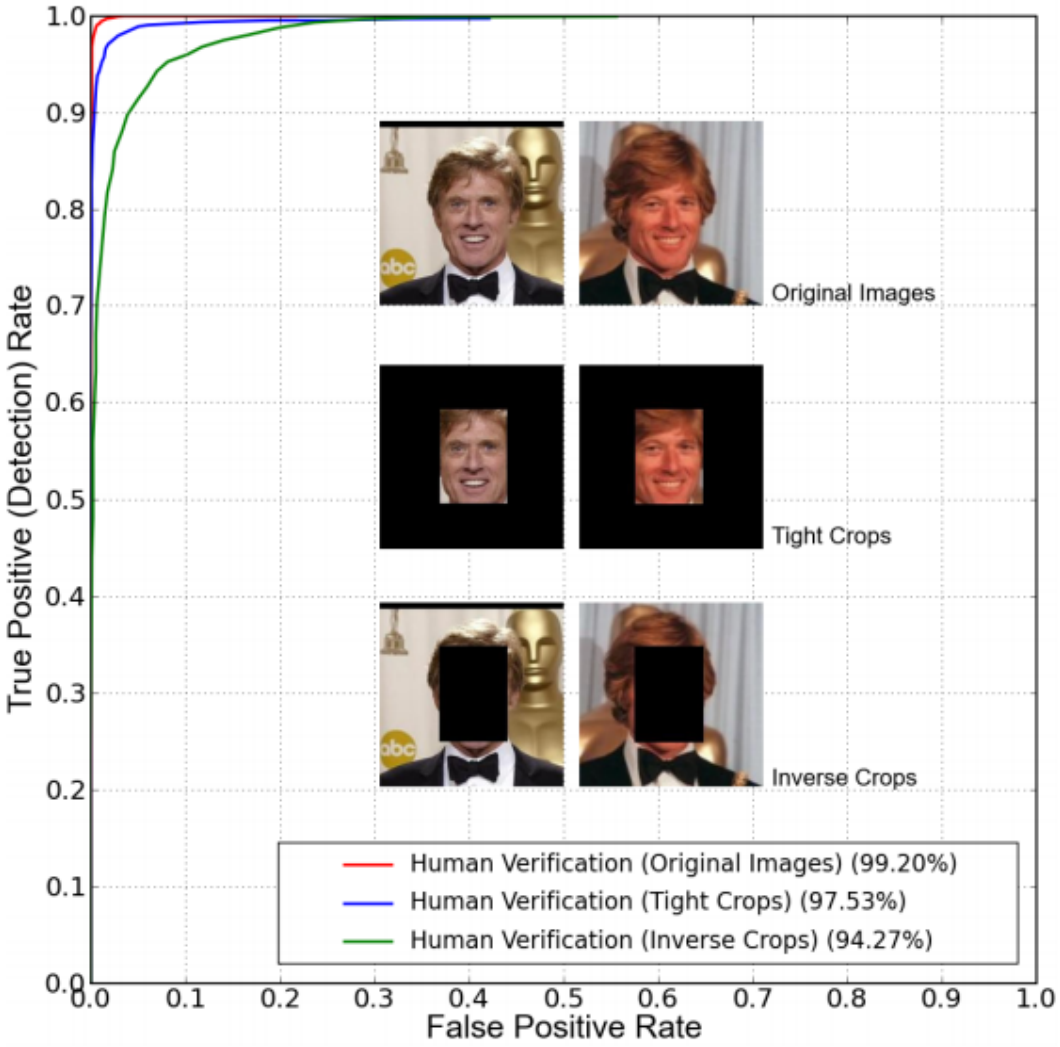


Figure 2.4: Performance of face recognition by humans on the LFW database: red line: using the original images, blue line: using a tighter cropped version of the images. The green line shows the performance of using inverse crop images (only the context is shown). Image is excerpted from [54].

hats (Figure 1.1c). The study showed that current commercial automatic face recognition systems have the notable potential to assist law enforcement. But the matching accuracy was not high enough and more progress must be made to increase the utility in face images taken in unconstrained environments.

Manual face recognition is highly robust to changes in pose and illumination, and also robust across the set of non-rigid face deformations such as expressions and speech movements. Although automatic face recognition techniques are widely used nowadays, there is still a notable gap between their performance and that of humans for face recognition in uncontrollable conditions. In the next section we will briefly introduce the face recognition algorithms for handling variation in illumination, pose, expression, low image quality such as blur and low-resolution, and ageing.

2.3 Unconstrained face recognition

Face recognitions system in real-world environments often have to contend with uncontrollable and unpredictable conditions such as large changes in illumination, pose, expression and occlusions, which introduce more intra-class variations and degrade the recognition performance. A large number of works have been proposed to deal with these variations. In this section we will briefly introduce the approaches for handling different variations and in the next section we will give a detailed review on the occlusion problem which is the main theme of this thesis. Although we do consider other unconstrained variations such as illumination and expression changes when dealing with the occlusion problem, we do not directly work on these variations such as ageing. A systematic survey of automatic face recognition before 2003 can be found in the work of [61]. The progression of recent automatic face recognition is presented in [62]. In this thesis, we focus on the 2D image based methods. The introduction to 3D face recognition techniques can be found in [63].

2.3.1 Approaches for handling illumination variations

Due to the 3D structure and various surface reflectance of faces, light sources can cast shading and shadows which accentuate or diminish certain facial features, creating non-uniform illumination on faces. The differences induced by these impact in the facial appearance can be greater than that between individuals. As analysed in [64], there are two categories of approaches for addressing the illumination problem - *active* and *passive* ways. The active approaches attempt to obtain face images which are invariant to the illumination changes. Usually specific devices such as 3D scanners, thermal cameras, infrared cameras, etc. other than the visible light cameras are required. The 3D shapes and reflectance of the facial surfaces, which are included in the 3D data, are the intrinsic factors of faces. 3D faces are more robust to the illumination changes compared with the 2D images. But they also incur higher computational cost for recognition. A good survey on 3D face recognition can be found in [63]. Thermal images and near-infrared images are more insensitive to large illumination changes as compared to visible light images. An introduction to illumination invariant face recognition using thermal images and near-infrared images is presented in [65] and [66], respectively.

On the other hand, passive approaches attempt to directly deal with images which have already been affected by illuminations. There are usually three classes of approaches: 1) illumination normalisation [67–69] which seeks to suppress the illumination variations either by image transformations or by synthesising an unaffected image from affected ones, 2) illumination invariant representation [50, 70] which attempts to extract features invariant to illumination changes, and 3) illumination variation modelling [25, 47, 71, 72] which is based on the theoretical principle that the set of images of a convex Lambertian object [73] obtained in a wide variety of illumination conditions can be approximated by a low-dimensional linear subspace in which the recognition can be performed [74].

2.3.2 Approaches for handling pose variations

Human beings are able to recognise a person by the face from different views. However, due to the complex 3D structures of faces, the view generalisation is still a tough task for automatic face recognition systems. A survey presented in [75] summarised the works that dealt with the pose problem in the past decades. The main challenge of pose problem is that appearance variations caused by variable poses can be larger than those caused by identity differences. The similarity between two faces from different persons in the same pose can be larger than that of the same person in different poses.

For pose-invariant face recognition based on 2D images, a natural solution is to collect multiple gallery images of all possible poses so the variations in the probe images can be covered. Multi-view methods [76, 77] are based on this requirement, which match a probe image with an arbitrary pose exhaustively against all gallery images and assign it to the class of the gallery image with the closest appearance. The multi-view methods are straight-forward and the frontal face recognition algorithms can be directly extended to solve the non-frontal pose problem. But the requirement of collecting sufficient gallery images limits the applicability of these algorithms. An alternative way is to synthesise different views of the probe face from a limited number of gallery images [78–80]. The synthesised face can be used as a bridge (i.e., a prototype face) to reduce the pose variations between the probe and the gallery. In recent years, patch-based methods have became popular [81–83]. A face is divided into patches and then the modelling of each corresponding patch pair is performed across poses. Local patches of a face are considered more robust to the pose variations compared with the holistic images/models. One limitation for pure 2D image based methods is that they assume the pose transformation is continuous within the 2D space. So usually this category of approaches is only able to handle small pose variations. On the other hand, approaches with assistance of 3D models [84–87] achieve better performance when addressing pose variations. Compared with 2D image based methods, 3D model based methods usually incur a high computational cost. Besides that, pose estimation is a sub-problem which is helpful in face recognition across

pose variations. Pose estimation is the process of inferring the orientation of a head. An systematic survey on head pose estimation is presented in [88].

2.3.3 Approaches for handling expression variations

Compared with pose and illumination variations, facial appearance changes caused by expressions are more *local* but more *dynamic*. Considering that expressions usually cause local distortions of a face, one idea for handling the expression problem is to use only regions less susceptible to expression variations. For example, one can just remove the mouth region for recognition since it changes largely due to expressions. But this *ad hoc* strategy generalises poorly to the general, realistic face images. Similar to the multi-view methods for handling pose variations mentioned in Section 2.3.2, another simple way to handle expression variations is to collect different expressions of each subject's face as the gallery. But the expression variations are more complicated. Because of the diversity of all possible expressions, there are a huge number of face images needed to be collected. Even though the expressions in the probe may still be substantially different from those in the gallery since expression changes are non-rigid and difficult to model.

Currently, there are two main categorises of approaches for handling expression variations: 1) weighting based methods [89, 90] and 2) warping based methods [91, 92]. Weighting based methods give lower weights to the pixels with large changes when matching the gallery (e.g., neutral faces) and the probe (e.g., smile faces). The optical flow [89] or Zernike moments [90] between two faces are used as a measure of the movements and texture changes due to expressions. Warping based methods try to warp an affected face to a neutral one, or synthesise a face with expressions from a neutral one. The morphable model [91] or optical flow [92] are used to estimate the motions of expressions. But not all face images can be well warped due to the lack of texture when the images contains very large deformations such as closed eyes and open mouths. Besides these approaches, 3D faces also have been used for expression-invariant recognition [93]. A comparative study of 3D face recognition against expression variations is presented in [94].

2.3.4 Approaches for handling low image quality variations

In security-related face recognition applications such as surveillance, usually the face images captured are degraded by low resolution and blur effects. For the low resolution problem, an intuitive solution is the super-resolution (SR) based method [95–99]. SR is a technique for synthesising high-resolution images from low resolution images for visual enhancement. After applying SR, a higher resolution image can be obtained and then used for recognition. One common drawback for SR based face recognition approaches is that SR does not directly contribute to recognition. The identity information may be contaminated by some artifacts attributed to the SR process. Another category of approaches do not apply the SR preprocessing to low resolution images. Popular approaches of this category include: support vector data description (SVDD) [100], coupled mappings (CMs) [101], multi-dimensional scaling [102], class specific dictionary learning [103], etc.

These are two types of effect attributed to the blur problem: focus blur and motion blur. A focus is the point where lights originating from a point on the object converge. When the light from object points is not well converged, a out-of-focus image will be generated by the sensor (e.g., camera) and results in the blur effect. The work in [104] analysed the impact of out-of-focus blur on face recognition performance. On the other hand, motion blur is due to the rapidly object movement or camera shaking. There are two main categories of approaches to improve the quality of the blurred face images: 1) blurred image modelling through subspace analysis [105] or sparse representation [106], and 2) blur-tolerant descriptors which attempt to extract blur insensitive features such as Local Phase Quantization (LPQ) [107, 108] to represent the face images.

2.3.5 Approaches for handling ageing

The typical application scenario of face recognition systems against the ageing effect is to detect if a particular person is present in a previous recorded database (e.g., missing children identification and suspect watch-list check). As the age between a probe and a

Figure 2.5: Face samples of the same individual across ages with appearance variations.

gallery image of the same subject increases, the accuracy of recognition system generally decreases. Ageing is a complex process that affects both the shape and texture (e.g., skin tone or wrinkles) of a face. From childhood to adulthood, the impact of ageing on facial appearance is mainly shape changes. After adulthood, the impact of ageing is mostly texture changes.

The automatic face recognition community has been taking two lines of investigation into the minimisation of the ageing effect on human faces: 1) developing *age estimation techniques* to classify face images [109, 110] and 2) developing *ageing insensitive algorithms* to perform recognition. In the early time, researchers tried to simulate the ageing effects by developing the ageing function and then performing automatic age estimation based on that [111]. But modelling the complex shape or texture variations of a face across ages is a very challenging task. Nowadays, researchers propose the generative ageing model [112] which learns a parametric aging model in the 3D domain to generate synthetic images and reduce the age gap between the probe and gallery images. One most challenging aspect of face recognition across ages is that it must address the variations of

all other aforementioned unconstrained factors as well because those factors also vary as time progresses.

Figure 2.5 shows the face samples of the same individual across different ages. Pose, expression, illumination changes occur when images are taken years apart, which give rise to large appearance changes of the faces. Besides the above unconstrained variations, the performance of face recognition can also be significantly affected by occlusions. The occlusion problem is relatively less studied in the automatic face recognition community. We will discuss this problem in details in the next section.

2.4 Face recognition with occlusions

The difficulty of occluded face recognition is twofold. First, occlusions distort the discriminative facial features and therefore increases the distance between two face images of the same subject in the feature space. As shown in Figure 2.6, the occluded face is actually from the *dot* class. However, since the eyes are occluded, it looks more similar to the faces from the *triangle* class, which will leads to misclassification. Holistic representation based methods such as Eigenface [38] and Fisherface [39] suffer a significant performance drop since the extracted features are related to all original pixels and can be easily affected even the occlusions are local. Occlusion errors on the original pixels become errors in the feature space and may even become less local.

The second difficulty is, face registration through the alignment of facial landmarks (e.g., the centres of the eyes, the tip of the nose) is needed before performing feature extraction. When facial landmarks are occluded, large registration errors usually occur and degrade the recognition rate. The study in [3] showed that the performance of face recognition approaches rely on the alignment accuracy. Different registration errors lead to different representations of the same face which can cause recognition failures [2].

As we mentioned in Chapter 1, some approaches first segment the occluded regions from face images and then perform recognition based on the remaining parts. Min *et*

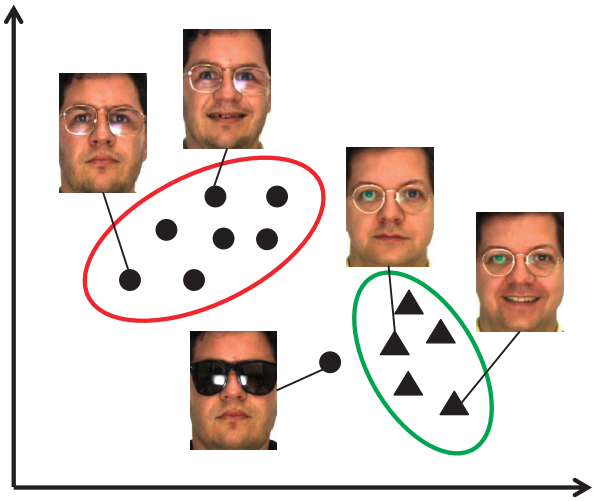


Figure 2.6: Occlusion problem: occlusions lead to a large distance between faces of the same person. The dots and triangles indicate two different classes. The occluded face belongs to the dot class.

al. [14] trained a SVM classifier to detect the occluded regions in a face image. Jia and Martínez modelled the occluded face recognition as a weighted linear least squares problem treating the occluded pixels as *missing data* (Partial within-class match (PWCM$_r$) [16], Partial support vector machines (PSVM) [17]). Their models require a skin colour based *occlusion mask* to detect the occluded areas in faces first. As we analysed before, the occlusion detectors are usually trained on specific types of occlusions (i.e., the training is data-dependent) and hence generalise poorly to various types of occlusions in real-world environments. So in this thesis we focus on performing recognition without detecting the occlusions in advance. There are three main categories of approaches for face recognition in the presence of occlusions: 1) reconstruction based approaches, 2) local matching based approaches, and 3) occlusion-insensitive feature based approaches.

2.4.1 Reconstruction based approach

Reconstruction based approaches treat face recognition with occlusions as a reconstruction (i.e., recovery) problem. The sparse representation based classification (SRC) [25] proposed by Wright *et al.* is a representative example. More details are presented in Section

3.1 since our proposed method is based on this model. A clean image is reconstructed from an occluded probe image through a linear combination of gallery images and the basis vectors of an occlusion dictionary. Then the occluded image is assigned to the class with the *best reconstruction*. Here the *best reconstruction* means reconstruction with a minimal matching error when comparing the reconstructed image with the probe image. The classification is based on the hypothesis that the most accurate reconstruction for the probe image can be obtained when using the gallery images from the correct class. SRC requires pixel-level alignment of images, which limits its use in practical applications.

There are some works aiming at improving the SRC model for occluded face recognition. Yang and Zhang [113] applied the Gabor feature to the original SRC model, which is coined Gabor-feature based SRC (GSRC), to reduce the size of occlusion dictionary and improve the robustness to occlusions. Zhou *et al.* [114] used a Markov Random Fields (MRF) model (SRC-MRF) to estimate the contiguous occlusions, which has been proved to be effective in improving SRC's performance. However, the performance of the SRC-MRF still drops drastically when the probe images are not well aligned. Chen *et al.* [115] applied the low-rank matrix recovery to decompose the gallery matrix (where each column is an image vector) into a representative basis matrix of low rank and a sparse error matrix. Then the low-rank matrix was used for reconstructing the occluded probe images as in the original SRC. This method is able to handle the occluded images in the gallery set but still requires the gallery images to be well aligned in advance.

Besides those, there are also other methods which follow the similar *reconstruction* idea but model the face representation differently. Naseem [116] modelled face recognition as a linear regression classification (LRC) problem. This method achieves competitive accuracy but is not robust to severe occlusions such as scarves. He *et al.* proposed the correntropy-based sparse representation (CESR) [117] with the maximum correntropy criterion, which can handle non-Gaussian noise (while SRC assumes that the reconstruction residuals follow the Gaussian or Laplacian distribution) and incur a lower computational cost than SRC. Similarly, Yang *et al.* proposed the robust sparse coding (RSC) [118],

which models the reconstruction as a sparsity constrained robust regression problem and is more robust to non-Gaussian noise and outliers (e.g., occlusions) than SRC. On the other hand, Zhang *et al.* [119] argued that the success of SRC actually comes from its collaborative representation over all classes of training samples and proposed the collaborative representation (CR) based classification with the regularised least square (CRC-RLS), which achieves competitive accuracy but with lower complexity than SRC.

A common drawback of reconstruction based methods, is that they usually requires a large number of samples per subject to represent a probe image. Most of them assume that the gallery/training images are captured in well controlled conditions. However, this assumption does not usually hold in real-world scenarios. Another drawback is this category of approaches usually incur a high computational cost [25].

2.4.2 Local matching based approach

With *local matching based approaches*, a face image is first partitioned into small regions (e.g., overlapping or non-overlapping patches, local facial components, small circles/ellipses), so the affected and unaffected parts of the face can be analysed in isolation. In order to minimise matching errors due to occluded parts, different matching strategies are performed.

Martínez proposed a local probabilistic method [2] which divides a face into k local parts then learns k eigenspaces from gallery images. Then all gallery face images are projected onto these subspaces and each class is modelled using k Gaussian mixture models (GMM). A probe image is also divided into k parts and projected onto the corresponding subspaces. Then the local distances (i.e., probabilities) to each GMM of a specific class can be computed and the sum of local distances can be used as the similarity measure between a given probe and a specific class. When a local area of the probe face is occluded, the occlusions only affect the local component. The mixture of Gaussian models accounts for the localisation errors of face images. Tan *et al.* [120] improved this method by using a self-organising map (SOM) instead of the mixture of Gaussians to model each class. The

advantage of using SOM is that compared with the mixture of Gaussians, it is unsupervised and nonparametric while still being able to eliminate possible noise and outliers. Moreover, based on SOM, Tan *et al.* [121] further applied the partial distance (PD) for occluded face recognition. It captures the significant partial similarities between face images, which is able to reduce the negative impacts of the unreliable features due to occlusions. Chen *et al.* proposed the Stringface [122, 123] which represents a face as a string (of line segments). So matching two faces is done by a *string-to-string* matching scheme, which is able to effectively find the most discriminative local parts without making any assumption on the distributions of the deformed facial regions affected by occlusions. But this method relies on the accuracy of line segmentation. Weng *et al.* [124] proposed a metric learned extended robust point matching (MLERPM) method based on the feature set matching. It considers the geometry information of feature sets and is robust to the misalignment of face images. But when dealing with randomly located occlusions, its performance degrades drastically.

The common intuition behind the local matching based approaches is that the facial appearance changes caused by occlusions are local in nature. Only parts of a face are distorted by occlusions while others are less affected and are reliable for recognition. So compared with the reconstruction based methods, local matching based methods are less likely to be able to handle the situation in which more than half of the face is occluded.

2.4.3 Occlusion-insensitive feature based approach

In addition to the above approaches, which focus on improving the robustness to occlusions during the matching stage, researchers also pay attention to image representation. The first category is related to the methods that encode the local features from only small areas of an image. Since they are locally computed, they are less likely to be affected by occlusions than the holistic features such as the Eigenface [38] and the Fisherface [39]. The local non-negative matrix factorization (LNMF) [125] and the locally salient independent component analysis (LS-ICA) [126], extract the local features using filter bases which are locally concentrated. Local image descriptors such as local Gabor binary pattern histogram

sequence (LGBPHS) [127], are also based on the *local* idea.

The second category of approaches attempt to extract occlusion-insensitive features from face images. Tzimiropoulos *et al.* [128] pointed out that PCA learning in the gradient orientation domain with a cosine-based distance measure helps reduce the effects due to occlusions in face images. The distribution of image gradient orientation (IGO) differences and the cosine kernel provide a powerful way to measure image correlation/similarity when image data are corrupted by occlusions. The face representations learned from the image gradient orientations are relatively invariant to the occlusion effects. Inspired by their work, Zhu *et al.* [129] further proposed a Gabor phase difference representation for occluded face recognition. They find that the Gabor phase (GP) difference is more stable and robust than gradient orientation to occlusions. Since the forms of occlusions in real-world scenarios are unpredictable, it is difficult to find a suitable representation which is insensitive to various types of occlusions.

Table 2.1 shows the categorisation of approaches in face recognition with occlusions. Note that the feature based approaches are not independent of the approaches of the two former categories. A sophisticated face recognition algorithm can combine both robust features and classifiers to handle the occlusion problem. For example, the MKD-SRC [5], which falls in the category of reconstruction based methods, involves the multi-task sparse representation learning. It extracted features from the local areas of a face and encoded these features using multi-keypoint descriptors (MKD) such as SIFT [130] and Gabor [131] based descriptor. These combined strategies make it robust to misalignment and occlusions in face images.

2.5 Challenges for occluded face recognition

2.5.1 The presence of occlusions

Most of the current methods assume that occlusions only exist in the probe images and the gallery or training images are *clean*. In real-world scenarios, the presence of occlusions is

Table 2.1: The categorisation of face recognition methods against occlusions

Recognition aided with occlusion detection	
SVM based occlusion detection [14]	
Partial withtin-class match (PWCM$_r$) [16]	
Partial support vector machines (PSVM) [17]	
Direct recognition in the presence of occlusions	
Reconstruction based	Sparse representation based classification (SRC) [25]
	Gabor-feature based SRC (GSRC) [113]
	SRC with Markov random field model (SRC-MRF) [114]
	Low-rank recovery [115]
	Linear regression classification (LRC) [116]
	Correntropy-based sparse representation (CESR) [117]
	Robust sparse coding (RSC) [118]
	Collaborative representation based classification (CRC-RLS) [119]
	Multi-keypoint descriptors based SRC (MKD-SRC) [5]
Local matching based	Local probabilistic subspace [2]
	Self-organising map (SOM) [120]
	Partial distance (PD) [121]
	Stringface [122, 123]
	Metric learned extended robust point matching (MLERPM) [124]
Occlusion-insensitive feature based	Local non-negative matrix factorization (LNMF) [125]
	Locally salient independent component analysis (LS-ICA) [126]
	Local Gabor binary pattern histogram sequence (LGBPHS) [127]
	Image gradient orientation based PCA (IGO-PCA) [128]
	Gabor phase (GP) [129]

unpredictable. Occlusions may occur in both gallery and probe images. We summarise in Table 2.2 three occlusion cases a face recognition system may encounter in the real-world applications. Most of the current methods rely on a clean gallery or training set and only consider the first case while the other two cases have not yet received much attention. As we analysed in Section 2.4, Jia and Martínez proposed a reconstruction based method [16], as well as an improved SVM [17] to handle occlusions in the training set. But the methods depend on a preprocessing of occlusion detection through the use of skin colour. Chen *et al.*'s work [115] uses the low-rank matrix recovery to deal with this problem. However, it requires faces to be well registered in advance. We will introduce our proposed algorithms which are able to deal with occlusions in both gallery and probe sets in the following chapters.

Table 2.2: Three typical occlusion cases

	Gallery	Probe	Scenario	Application
Uvs.O:	Unoccluded	Occluded	Access control, ID management	Facility access, immigration, user registration, online authentication
Ovs.U:	Occluded	Unoccluded	Law enforcement, security, surveillance	Suspect investigation, shoplifter recognition, criminal face retrieval, terrorist alert, CCTV control
Ovs.O:	Occluded	Occluded		

2.5.2 The prior knowledge of occlusions

In the real-world scenarios, there is almost no prior knowledge of occlusions. The type, location, area, texture and shape of occlusions are unpredictable. Some approaches assume that occlusions either occur on the upper half of the face (e.g., sunglasses) or the lower half (e.g., scarf). This prior knowledge is helpful for the recognition on a specific testing dataset (e.g., the AR database [132]) but generalise poorly to general occlusions in real-world environments. When designing a robust algorithm, specific assumptions about the aforementioned factors should be avoided. On the other hand, in real-world environments, occlusions are usually coupled with other factors (e.g., illumination, expression changes). So when conducting experiments to evaluate the effectiveness of an occluded face recognition algorithm, at least three kinds of occluded images should be considered:

1. Images containing real occlusions such as disguises with various textures and shapes.
2. Images containing randomly located occlusions without any prior knowledge of the location.
3. Images taken in natural conditions (e.g., outdoor environments) in which occlusions are coupled with other distorting factors.

These three types of images listed in Table 2.3 are usually partly or even fully ignored by existing works. In this thesis we consider all types of occluded face images in the experiments. Details about the databases will be introduced in the next section.

Table 2.3: Three types of occluded face images

Type	Factor	Available database
With natural occlusions	Various textures and shapes	AR [132]
With randomly located occlusions	No location prior knowledge	Synthetic occlusions on the standard databases such as Extended Yale B [133], FRGC [52]
Natural images	Occlusions coupled with other variations	LFW [134], TFWM [135]

2.6 Databases

Over the past four decades, a large number of face databases have been built to evaluate the effectiveness of face recognition algorithms. Especially for the PIE problems, specific databases have been designed, which promote the development of face recognition algorithms for these factors. However, there are very few databases to support research into the occlusion related problems. As analysed in Section 2.5.2, at least three types of occluded face images should be considered when evaluating the effectiveness of face recognition algorithms. In this section we will introduce the databases according to the three types of occluded images as shown in Table 2.3.

2.6.1 The AR database

The AR database is one of the very few databases which contain real occlusions and are open to the public. It contains over 4,000 colour images of 126 subjects' faces (70 men and 56 women). These images (Figure 2.7) suffer from different variations in facial expressions, illumination conditions and occlusions (i.e., sunglasses and scarves). They were taken under strictly controlled conditions. No restrictions on wear (clothes, glasses, etc.), make-up, hair style, etc. were imposed to subjects. For each subject, 26 images in total were taken in two sessions (two weeks apart). For each session (13 images), the descriptions of images are as follows:

1. Neutral expression

2. Smile
3. Anger
4. Scream
5. Left light on
6. Right light on
7. All side lights on
8. Wearing sun glasses
9. Wearing sun glasses and left light on
10. Wearing sun glasses and right light on
11. Wearing scarf
12. Wearing scarf and left light on
13. Wearing scarf and right light on

The limitations of the AR database are that it only contains two types of occlusions, i.e., sunglasses and scarf, and the location of the occlusion is either on the upper face or lower face. As a result, researchers also simulate occlusions (e.g., randomly located occlusions) using images from the standard databases such as the Extended Yale B database [133] and the FRGC database [52] to evaluate their proposed methods. The schemes for generating the synthetic occlusions vary across works, so in the next sub-sections we will only introduce the most popular one which is well accepted by the public.

2.6.2 The Extended Yale B database

Extended Yale B database [133] contains 2,414 frontal face images of 38 persons in 64 different illumination conditions. For every subject in a particular pose, an image with

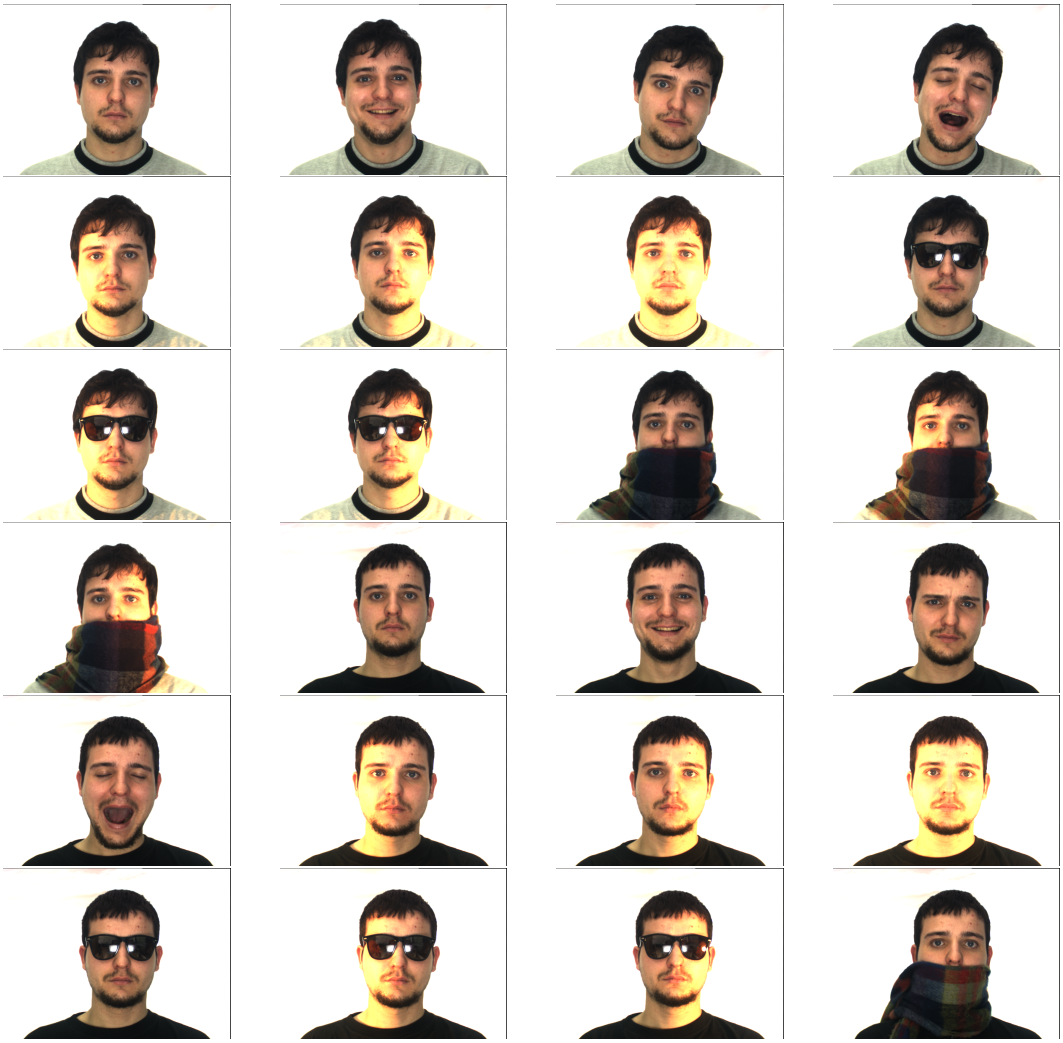


Figure 2.7: Sample images of two sessions from the AR database.

ambient (background) illumination was also captured. The images are grouped into four subsets according to the lighting angle with respect to the camera axis. The Subset 1 and Subset 2 cover the angular range 0° to 25°, the Subset 3 covers 25° to 50°, the Subset 4 covers 50° to 77°, and the Subset 5 covers angles which are larger than 78°.

In order to simulate different levels of contiguous occlusions, the most used scheme [25] is to replace a randomly located square patch from each test image with a baboon image which has similar texture with the human face. The location of the occlusion is randomly chosen. The sizes of the synthetic occlusions vary in the range of 10% to 50% of the original image (Figure 2.8).

2.6.3 The FRGC database

The Face Recognition Grand Challenge (FRGC) database [52] contains 8,014 images from 466 subjects in difference sessions. For each subject in each session, there are four controlled still images, two uncontrolled still images, and one 3D image. In our experiments we only use the still images. These images contain variations such as illumination and expression changes, time-lapse, etc. The controlled images were full frontal facial images taken under two lighting conditions (two or three studio lights) and with two facial expressions (smiling and neutral). The uncontrolled images were taken in varying illumination conditions; e.g., hallways, atria, or outdoors. Each set of uncontrolled images contains two expressions, smiling and neutral.

To simulate the randomly located occlusions, one can replace a randomly located square patch from each image with a black block as discussed in [17]. The location of the occlusion is randomly chosen. The size of the black block varies in the range of 10% to 50% of the original image (Figure 2.9).

2.6.4 The LFW database

The *Labeled Faces in the Wild* (LFW) database [134] is a database of face photographs designed for studying the problem of unconstrained face recognition. As mentioned in

(a) Subset 1

(b) Subset 2

(c) Subset 3

(d) Subset 4

(e) Subset 5

Figure 2.8: Sample images from the Extended Yale B database with randomly located occlusions: a) Subset 1, b) Subset 2. c) Subset 3, d) Subset 4 and e) Subset 5.

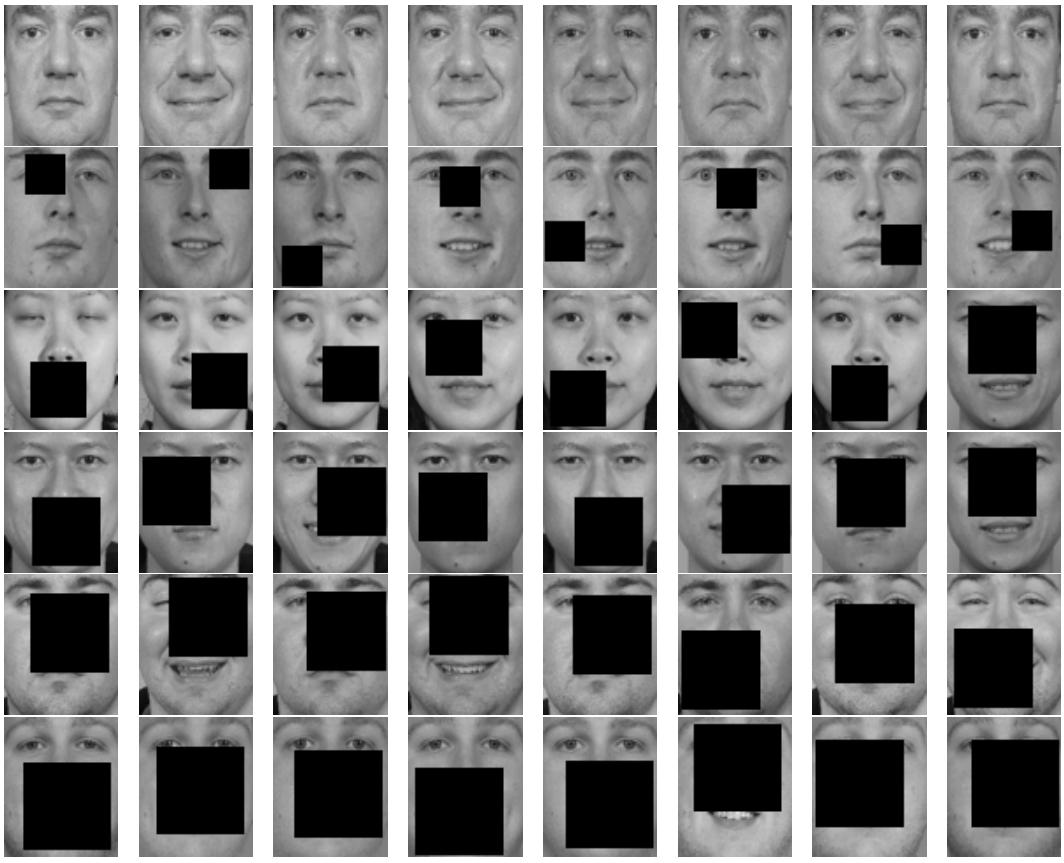


Figure 2.9: Sample images from the FRGC database with randomly located occlusions.

Figure 2.10: Sample images from the LFW database: first and second row: six matched pairs from six subjects, and third and forth row: six non-matched pairs from twelve subjects.

Section 2.2.2, it contains 13,233 face images of 5,749 subjects collected from the Internet. These images are captured in uncontrollable environments and contain large variations in pose, illumination, expression, time-lapse and various types of occlusions. The only constraint on these faces is that they were detected by the Viola-Jones face detector [55]. Each face has been labelled with the name of the subject pictured. 1,680 of the subjects pictured have two or more distinct images in the database. The task of face verification under the LFW database's protocol is to determine if a pair of face images belongs to the same subject or not (Figure 2.10).

2.6.5 The TFWM database

In this thesis, we also introduce a database called *The face we make* (TFWM) [135] collected by a designer Dexter Miranda. We have got permission from the author for using these images for research purposes. This database is very suitable for evaluating the occluded face recognition algorithms, as well as methods in other face related problems.

Figure 2.11: Sample images from the TFWM database.

The database has more than 2,000 images of more than 230 subjects. It contains frontal view faces of strangers on the streets with uncontrollable lighting. The sources of occlusions include glasses, sunglasses, hat, hair and hand on the face. These images are captured in natural and arbitrary conditions. Besides occlusions, these images also contain expression, pose and head rotation variations (Figure 2.11).

2.7 Summary

In this chapter, we first introduced face recognition approaches performed by both human beings and computers. We then discussed the difficulties of unconstrained face recognition and further introduced the occlusion problem which is the main theme of this thesis. We have introduced the approaches for occluded face recognition and summarised the current challenges. Finally we introduced the popular databases used for the research of occlusion

problems in face recognition. We will present our proposed methods for occluded face recognition in the following chapters.

Chapter 3

Structured sparse representation (SSR) based face recognition

In this chapter, we introduce the reconstruction based method to deal with occlusion related problems. As mentioned in Chapter 2, sparse representation based classification (SRC) [25] represents a probe image as a linear combination of a minimal number of gallery images. In this chapter, we propose the structured sparse reconstruction (SSR) based approach considering the *structured sparsity* [136]. Face images from the same class/subject span a sub-space in a high-dimensional feature space. These images from the same class can be seen as a *cluster* in the whole gallery set. In the SSR model, a probe image is represented as a linear combination of a minimal number of clusters. Compared with SRC which represents an image using gallery images from all classes, SSR represents an image by just involving the images from the most probable classes, which is more suitable for classification. To better model occlusions, a structured occlusion dictionary is proposed. Our work points out that the non-zeros entries in the occlusion coefficient vector also have a cluster structure because occlusions appear in regions in an image and the occluded pixels are spatially contiguous.

In real-world environments, occlusions and other variable factors usually coexist. Dealing with multiple occlusions and other variable factors is very practical, but difficult.

In this chapter we consider the coupled condition of extreme illuminations and occlusions for face recognition. There are two observations for doing so: 1) strong shadows under extreme illuminations can be seen as partial occlusions on the face, and 2) as pointed out by the work in [47], face images in an illumination condition can be reconstructed from a linear combination of face images taken with different lighting directions. So the illumination and occlusion problems can be solved in the same reconstruction based framework. We adopt an illumination insensitive Weber local descriptor (WLD) [26] in the SSR model. It not only makes our method strongly resist the illumination changes, but also eliminates the shadows that are used to be modelled as sparse error. All of these can reduce the representation errors and help to produce better classification results.

The rest of this chapter is organised as follows. Section 3.1 briefly introduces the SRC model. Section 3.2 explains the proposed approach in two steps: 1) the SSR and 2) the structured occlusion dictionary. And then, Section 3.3 describes the strategy of combining the SSR and the robust WLD [26]. To evaluate the effectiveness of the proposed approach, extensive experiments are conducted and the results are reported in Section 3.4. Finally, Section 3.5 concludes this chapter. Note that symbols used are only valid within this chapter.

3.1 Sparse representation based classification

Inspired by the findings that natural images can be generally represented by structural primitives (e.g., edges and line segments) that are qualitatively similar in form to simple cell receptive fields [137], the sparse representation encodes a signal using a small number of atoms from an over-complete dictionary. Here an over-complete dictionary is a collection of base elements (signals) where the number of elements exceeds the dimension of the signal. In this way any signal can be represented by more than one combination of different bases. The sparse representation has attracted lots of attention in the image processing community [138, 139]. A review of sparse representation for computer vision and pattern recognition

applications is presented in [140].

The sparse representation based classification (SRC) for face recognition proposed in [25] represents a probe face image in an over-complete dictionary whose base elements are the gallery face images themselves. SRC assumes that sufficient gallery samples are available for each class so the probe sample can be well represented by a linear combination of the gallery samples from the same class. Then the probe sample will be classified as the class with the minimal representation residual.

Arranging a set of n_i gallery images from i-th class as columns of a matrix $\boldsymbol{X}_i = [\boldsymbol{x}_{i,1}, ..., \boldsymbol{x}_{i,n_i}]$, the whole gallery set $\boldsymbol{X} \in \mathbb{R}^{m \times n}$ is the concatenation of n gallery images from all C classes:

$$\boldsymbol{X} = [\boldsymbol{X}_1, ..., \boldsymbol{X}_i, ..., \boldsymbol{X}_C] = [\boldsymbol{x}_{1,1}, ..., \boldsymbol{x}_{1,n_1}, ..., \boldsymbol{x}_{i,1}, ..., \boldsymbol{x}_{i,n_i}, ..., \boldsymbol{x}_{C,1}, ..., \boldsymbol{x}_{C,n_C}] \quad (3.1)$$

where each column of $\boldsymbol{X}$ is an image vector of length m and $n = n_1 + ... + n_i + ... + n_C$. A probe image $\boldsymbol{y} \in \mathbb{R}^m$ is represented as a linear combination of gallery images:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\alpha} \quad (3.2)$$

where $\boldsymbol{\alpha} \in \mathbb{R}^n$ is a coefficient vector whose entries are zero[1] except those associated with the gallery images from the class of $\boldsymbol{y}$. It can be similarly denoted as:

$$\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1, ..., \boldsymbol{\alpha}_i, ..., \boldsymbol{\alpha}_C] = [\alpha_{1,1}, ..., \alpha_{1,n_1}, ..., \alpha_{i,1}, ..., \alpha_{i,n_i}, ..., \alpha_{C,1}, ..., \alpha_{C,n_C}] \quad (3.3)$$

where $\boldsymbol{\alpha}_i$ is the set of coefficients associated with the i-th class.

Generally, a probe image can be well represented only by the images from its class, so the coefficient vector $\boldsymbol{\alpha}$ just contains a small number of non-zero entries. Thus, this sparse vector $\boldsymbol{\alpha}$ encodes the identity of $\boldsymbol{y}$ which can be obtained by solving the following

[1] In fact, in the implementation, these entries are not always zero, but rather relatively small in magnitude.

optimisation problem:

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\alpha} \|\boldsymbol{\alpha}\|_0 \quad \text{subject to} \quad \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\alpha} \tag{3.4}$$

where $\|\cdot\|_0$ is the l_0-norm which corresponds to the number of non-zero elements in a vector. Assuming that the underlying subspace for each class is low-dimensional and the number of gallery images is sufficient (i.e., $m < n$), the system of equations $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\alpha}$ is underdetermined. The problem (Equation 3.4) of finding the sparsest solution of such a underdetermined system of linear equations is NP-hard [141]. Fortunately, the theory of sparse representation and compressed sensing reveals that if $\boldsymbol{\alpha}$ is sparse enough, the solution of the l_0-minimisation problem in (3.4) can be obtained by replacing the l_0-norm with l_1-norm [142, 143]:

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\alpha} \|\boldsymbol{\alpha}\|_1 \quad \text{subject to} \quad \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\alpha} \tag{3.5}$$

where $\|\cdot\|_1$ indicates the l_1-norm which is the sum of the absolute value for each element in a vector.

Considering the occlusions in a face image, an occluded probe image $\boldsymbol{y}'$ is represented using the linear model modified from (3.2) as:

$$\boldsymbol{y}' = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{e} \tag{3.6}$$

where and $\boldsymbol{e} \in \mathbb{R}^m$ is the occlusion error. For implementation, (3.6) is rewritten as:

$$\boldsymbol{y}' = [\ \boldsymbol{X} \quad \boldsymbol{I}\ ][\ \begin{matrix}\boldsymbol{\alpha} \\ \boldsymbol{\alpha_e}\end{matrix}\ ] = \boldsymbol{B}\boldsymbol{\omega} \tag{3.7}$$

where $\boldsymbol{e} = \boldsymbol{I}\boldsymbol{\alpha_e} \in \mathbb{R}^m$ and $\boldsymbol{\alpha_e}$ is the corresponding coefficient vector for the occlusions. $\boldsymbol{I} \in \mathbb{R}^{m \times m}$ is an identity matrix which can be called the *occlusion dictionary*. Each column of $\boldsymbol{I} \in \mathbb{R}^{m \times m}$ can be seen as a black image with only one white pixel. So an identity matrix

of size $m \times m$ is able to represent the pixel of occlusions appearing in arbitrary location in an image with the dimension of m (i.e., $m = w \times h$ where w and h are the width and height of an image). $\boldsymbol{B} = [\boldsymbol{X} \quad \boldsymbol{I}] \in \mathbb{R}^{m \times (n+m)}$ is the concatenation dictionary which is a concatenated matrix of $\boldsymbol{X}$ and $\boldsymbol{I}$. $\boldsymbol{\omega} = [\boldsymbol{\alpha}; \boldsymbol{\alpha_e}] \in \mathbb{R}^{n+m}$ is the concatenation coefficient vector which is a concatenated column vector of $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha_e}$. In this form, the optimisation problem in (3.5) is extended to:

$$\hat{\boldsymbol{\omega}} = \arg\min_{\omega} \|\boldsymbol{\omega}\|_1 \quad \text{subject to} \quad \boldsymbol{y'} = \boldsymbol{B\omega} \tag{3.8}$$

where $\hat{\boldsymbol{\omega}} = [\hat{\boldsymbol{\alpha}}; \hat{\boldsymbol{\alpha}}_e]$ encodes the identity of $\boldsymbol{y'}$ while considering the occlusion effect.

Finally, the probe image $\boldsymbol{y'}$ is assigned to the class which minimises the reconstruction error:

$$class(\boldsymbol{y'}) = \arg\min_{i} \left\|\boldsymbol{y'} - \boldsymbol{X}_i\hat{\boldsymbol{\alpha}}_i - \boldsymbol{I}\hat{\boldsymbol{\alpha}}_{\boldsymbol{e}}\right\|_2 \tag{3.9}$$

where $\boldsymbol{X}_i\hat{\boldsymbol{\alpha}}_i$ can be seen as the reconstructed image recovered by the gallery images from the i-th class.

Up to now, we have introduced the SRC model. The structured sparse representation (SSR) model also follows the classification framework as SRC. However, it takes a different view for representing an image. We will explain the differences in the next section.

3.2 Structured sparse representation based face recognition

3.2.1 Structured sparse representation

SRC models a probe image by seeking the sparsest representation of the set of gallery images across all classes, which is optimal for *reconstruction* purposes, but not necessarily for *classification* tasks. Considering the example in Figure 3.1, a probe image (middle) of Class 1 can be modelled as a linear combination of either 1) one image from Class 1, one image from Class 2 and one image from Class 3, respectively (Figure 3.1a), or 2) three images from Class 1 (Figure 3.1b). These two representations are equivalent from a sparest

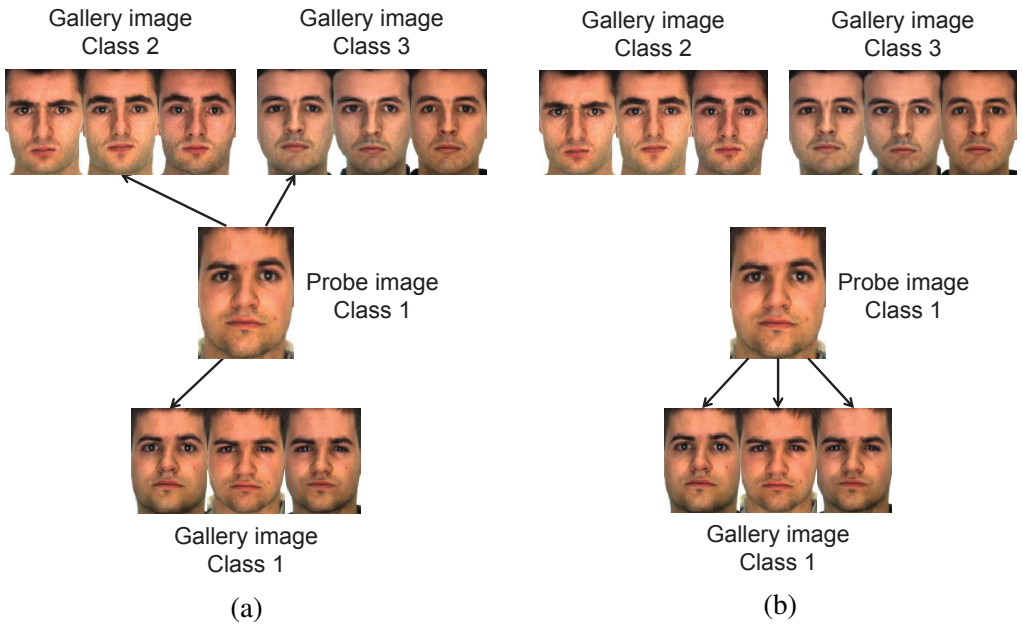


Figure 3.1: An illustration of the sparse representation. A probe image of Class 1 can be represented as a linear combination of a) one image from Class 1, one image from Class 2 and one image from Class 3; as well as b) a linear combination of three images from Class 3.

representation perspective because they use the same number of base images to represent an image. In classification, a probe image is classified by evaluating which class contributes the most for representing it. So from the classification point of view, the representation in Figure 3.1b is better for decision making.

In our approach, we model a face image through a structured sparse representation way. Instead of achieving the *flat sparsity* using the gallery images across all classes as in SRC, SSR achieves the *structured sparsity* by using a minimum number of classes attempting to choose the gallery images from the correct class. The concept of *structured sparsity* is also widely adopted in the computer vision applications such as visual tracking [144], object classification [145], image annotation [146], etc.

Face images from the same class can be seen as a *cluster* in a high-dimensional feature space. So the $\boldsymbol{X}$ in (3.1) is a concatenation of C clusters, where C is the number of classes and the i-th cluster contains n_i images. Correspondingly, the sparse coefficient

vector $\boldsymbol{\alpha}$ in (3.3) also has a cluster structure.

SSR tries to find the sparsest representation of a probe image using a minimal number of clusters where each cluster contains the gallery images from the same class. Thus, the representation of the probe image just involves the most probable classes, which is more suitable for classification purposes than SRC. Compared with the l_1-norm optimisation[2] in (3.5), the sparsest solution of SSR can be obtained through solving the following mixed l_2/l_1-norm optimisation problem:

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\alpha} \|\boldsymbol{\alpha}\|_{2,1} = \arg\min_{\alpha} \sum_{i=1}^{C} \sqrt{\sum_{j=1}^{n_i} \alpha_{ij}^2} \quad \text{subject to} \quad \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\alpha} \tag{3.10}$$

where l_2/l_1-norm [149] is a rotational invariant l_1-norm. The inner l_2-norm of $\|\boldsymbol{\alpha}\|_{2,1}$ enforces the selection of all the coefficients within a cluster while the outer l_1-norm promotes sparsity in the number of selected clusters. When the size of each cluster equals to 1 (i.e., each image in the gallery set is seen as a cluster), the structured sparsity reduces to the conventional sparsity in SRC.

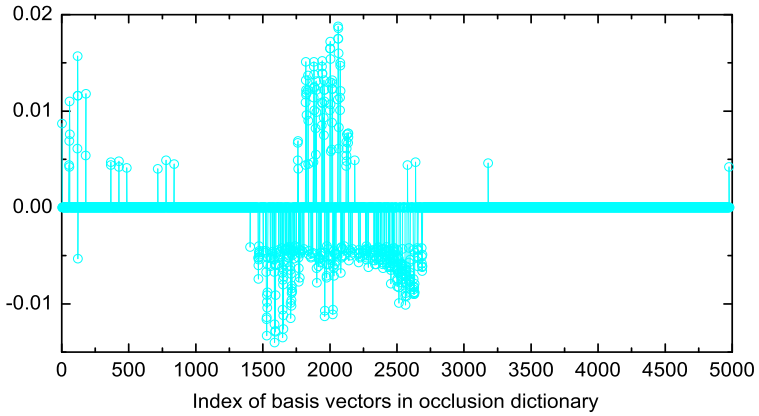
3.2.2 Structured occlusion dictionary

To deal with occlusions, an occluded image $\boldsymbol{y}' \in \mathbb{R}^m$ is approximated by a clean image $\boldsymbol{y} \in \mathbb{R}^m$ plus an error vector $\boldsymbol{e} \in \mathbb{R}^m$ as shown in (3.6). According to (3.7), the occlusion dictionary is set as the identity matrix $\boldsymbol{I} \in \mathbb{R}^{m \times m}$ [25, 150]. The occlusion error $\boldsymbol{e} \in \mathbb{R}^m$ is represented by a few of basis vectors of $\boldsymbol{I}$. However, an identity matrix $\boldsymbol{I} \in \mathbb{R}^{m \times m}$ is able to represent any image of size m without $\boldsymbol{X}$ in (3.7). Thus, face pixels may be represented by $\boldsymbol{I}$ and as a result incorrectly processed as occlusions. To accurately model the *real* occlusions, we propose a structured occlusion dictionary $\boldsymbol{D}$ to replace the identity matrix $\boldsymbol{I}$ in (3.7):

$$\boldsymbol{y}' = [\ \boldsymbol{X} \quad \boldsymbol{D}\ ][\ \begin{matrix}\boldsymbol{\alpha} \\ \boldsymbol{\alpha_e}\end{matrix}\ ] \tag{3.11}$$

[2] A discussion of the use of l_1-norm can be found in [147] and [148].

(a)



(b)

Figure 3.2: An illustration of the cluster structure of occlusion coefficients. a) A probe image with sunglasses occlusion. b) The corresponding occlusion coefficient. The entries with significant value are spatially contiguous and aligned as clusters.

So the problem now is how to set $\boldsymbol{D}$. Taking into account the spatial distribution of contiguous occlusions [114, 151], the non-zero entries in the occlusion coefficient vector $\boldsymbol{\alpha_e}$ are likely to be spatially continuous, that is, the non-zero entries are aligned to *clusters* instead of arbitrarily spread throughout the coefficient vector. An illustration is shown in Figure 3.2. The occlusion occurs in the upper face (Figure 3.2a) so entries with significant value are mainly distributed in the initial part of the coefficient vector (Figure 3.2b). This indicates that the occlusion dictionary $\boldsymbol{D}$ also has a cluster structure where the spatially contiguous bases form several clusters. Similarly to the $\boldsymbol{X}$ in (3.1), the occlusion dictionary $\boldsymbol{D}$ is also a concatenation of q clusters as:

$$\boldsymbol{D} = [\boldsymbol{D}_1, \boldsymbol{D}_2, ..., \boldsymbol{D}_q] = [\boldsymbol{a}_1, ..., \boldsymbol{a}_{d'}, \boldsymbol{a}_{d'+1}, ..., \boldsymbol{a}_{2d'}, ..., \boldsymbol{a}_{m-d'+1}, ..., \boldsymbol{a}_m] \tag{3.12}$$

where the j-th cluster $\boldsymbol{D}_j$ contains $d'(1 \leqslant d' \leqslant m)$ spatially contiguous columns of the identity matrix $\boldsymbol{I} \in \mathbb{R}^{m \times m}$, $\boldsymbol{a}_j$ is the j-th column of $\boldsymbol{I}$ and $m = d' \times q$.

Similar to (3.8), let $\boldsymbol{B} = [\boldsymbol{X} \quad \boldsymbol{D}] \in \mathbb{R}^{m \times (n+m)}, \boldsymbol{\omega} = [\boldsymbol{\alpha}; \boldsymbol{\alpha_e}] \in \mathbb{R}^{n+m}$ as the

concatenation dictionary and concatenation coefficient vector, respectively, the structured sparsest representation can be obtained within the same framework as in (3.10) as[3]:

$$\hat{\boldsymbol{\omega}} = \arg\min_{\omega} \|\boldsymbol{\omega}\|_{2,1} \quad \text{subject to} \quad \boldsymbol{y'} = \boldsymbol{B\omega} \tag{3.13}$$

In this framework, $\boldsymbol{D}$ only represents occlusions and $\boldsymbol{X}$ represents faces, which can help to reduce the representation error. This is guaranteed by seeking the sparsest representation using a minimal number of clusters within $\boldsymbol{X}$ and $\boldsymbol{D}$.

Finally, similar to SRC in (3.9), the probe image $\boldsymbol{y'}$ can be classified as the class with the smallest residual:

$$class(\boldsymbol{y'}) = \arg\min_{i} \left\|\boldsymbol{y'} - \boldsymbol{X}_i\hat{\boldsymbol{\alpha}}_i - \boldsymbol{D}\hat{\boldsymbol{\alpha}_e}\right\|_2 \tag{3.14}$$

Figure 3.3 illustrates the framework of the SSR model aided with the structured occlusion dictionary. An occluded image can be seen as a reconstructed image plus the occlusion error. It can be found that both the gallery set and the occlusion dictionary have a cluster structure. The probe image is represented as a linear combination of gallery images and occlusion bases using a minimal number of clusters. The whole process of the proposed method is shown in Algorithm 1.

Algorithm 1 Structured sparse representation based occluded face recognition algorithm

Input:

$\boldsymbol{B}$: a concatenation matrix of the gallery images $\boldsymbol{X}$ and the occlusion dictionary $\boldsymbol{D}$;
$\boldsymbol{y'}$: a probe image

Output:

class: the class label of $\boldsymbol{y'}$;

1: Normalise the each column of $\boldsymbol{B}$ to have a unit l_2-norm;
2: Solve the l_2/l_1-minimisation problem:
$\hat{\boldsymbol{\omega}} = \arg\min_{\omega} \|\boldsymbol{\omega}\|_{2,1}$ subject to $\boldsymbol{y'} = \boldsymbol{B\omega}$ where $\hat{\boldsymbol{\omega}} = [\hat{\boldsymbol{\alpha}}; \hat{\boldsymbol{\alpha}_e}]$;
3: Compute the residuals and output the label of the class with the smallest residual:
$class(\boldsymbol{y'}) = \arg\min_i \|\boldsymbol{y'} - \boldsymbol{X}_i\hat{\boldsymbol{\alpha}}_i - \boldsymbol{D}\hat{\boldsymbol{\alpha}_e}\|_2$;

[3] the optimisation problem can be solved by algorithms such as [152, 153]

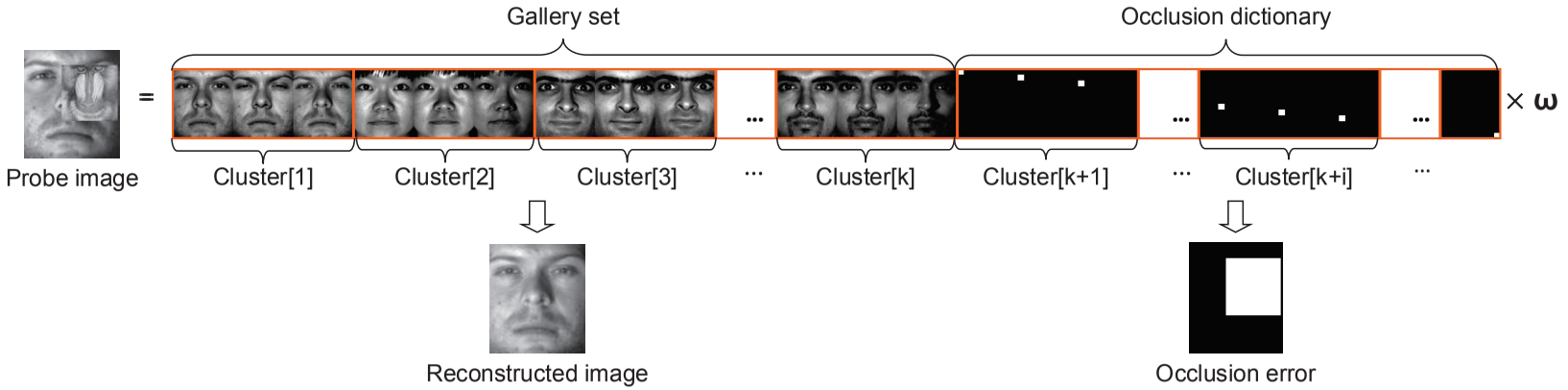


Figure 3.3: Structured sparse representation aided with the cluster occlusion dictionary. A probe image is represented as a linear combination of gallery images and occlusion bases using a minimal number of clusters. ω is the sparse coefficient vector. The gallery images of the same person form a cluster in the gallery set. Each cluster of the occlusion dictionary contains $d'(d' = 3$ in this case) spatial contiguous columns of the identity matrix $\boldsymbol{I}$ (each column is illustrated as a black image with only one white point).

3.3 Combining SSR with illumination insensitive features

In real-world environments, a probe image may lie outside the sub-space spanned by the gallery images due to extreme illumination conditions. Moreover, shadows plus other occlusions can distort or even obscure facial features. In order to handle the coupled effect of extreme illuminations and occlusions, we adopt a discriminative Weber local descriptor (WLD) [26] in the SSR model.

Inspired by the psychophysical *Weber's Law*[4] [154], WLD is compatible with the human visual perception. It has been successfully applied to the area of texture analysis [26]. We adopt the differential excitation component of WLD since it is insensitive to the illumination variations.

For a given pixel p, its WLD value $f(p)$ can be computed as a function of the ratio between two terms: the intensity of p and the relative intensity differences between p and its neighbours $p_i (i = 1, 2, ..., l, l$ is the number of neighbours) [26]:

$$f(p) = \arctan(\sum_{i=1}^{l} \frac{p_i - p}{p}) \tag{3.15}$$

Arctangent function is used here since it can limit the output to prevent it from increasing or decresing too quickly when the input becomes larger or smaller.

For each face image, the WLD features are computed pixel by pixel using (3.15) and then are converted to a single vector as a column of the gallery matrix $\boldsymbol{X}$ in (3.1). WLD calculates the intensity differences between a pixel and its neighbours and thus a brightness change of adding a constant to each pixel will not influence the differences value. In addition, the intensity difference is divided by the intensity of the given pixel, so a contrast change of multiplying each pixel value by a constant will be cancelled by the division. This indicates that WLD is robust to uniform illumination changes. Within a shaded area, the main effect on a small neighbourhood of a pixel (e.g., a 3×3 pixels patch, $l = 8$ used in this chapter) can be assumed to be uniform due to its compactness. Therefore,

[4]Weber's Law: *the ratio of the increment threshold to the background intensity is a constant*

the main effect of shadow can be corrected by WLD. An example of WLD feature is shown as Figure 3.4e and Figure 3.4f.

We adopt WLD in the SSR model for the following reasons: 1) WLD feature is computed pixel by pixel, each time within a small neighbourhood. On the one hand, this densely extracted feature is able to maintain detailed facial information that is effective for recognition. On the other hand, the WLD feature is a local feature, which is less likely to be corrupted by occlusions compared with the holistic features [38, 39]. 2) WLD is strongly robust to illumination changes. Images can be more easily represented by the existing gallery images. 3) The facial areas obscured by shadows are usually processed as sparse error. WLD is insensitive to shadows and the affected areas can be applied to recover the unoccluded image. In other words, more information can be used for recognition. There are also other illumination insensitive features [69, 155, 156] as well as other edge detection filters (eg, the Laplacian of Gaussian), but the results in [157] indicate that WLD works best so we adopt it in our model. More detailed discussions of the comparison between WLD and other features can be found in the original paper [26].

3.4 Experimental analysis

To evaluate the performance of the proposed model, a set of large-scale identification experiments are conducted in this section. We test our model and other sparse representation based methods which have been proved to be effective for robust face recognition in the literature:

1. SRC-I [25]: the original SRC using the pixel intensity.
2. SRC-W: the SRC using the WLD features.
3. SRC-G [113]: the SRC using the Gabor features. In our experiments, we follow the same implementation setting as the work in [113].
4. SSR-I: the SSR using the pixel intensity values.

5. SSR-W: the SSR using the WLD features.

We first evaluate the proposed approach using images from the Extended Yale B database [47] with randomly located occlusions in extreme illumination conditions. We compare the results of using and not using WLD features and give an example to explain how WLD features improves the performance of sparse representation based methods. Then we test the SSR model on the AR database [132] with both real disguise and non-uniform illuminations. The effect of cluster size is also discussed. Note that in all experiments, the gallery image set is disjoint with all probe sets. The cluster size in the cluster dictionary is set to 20 (the effect of cluster size is discussed in Section 3.4.3).

3.4.1 Face identification with randomly located occlusions and extreme illuminations

We first test our approaches on the Extended Yale B database [133]. In order to simulate different levels (from 0% to 50%) of contiguous occlusions, we replace a randomly located square patch from each test image with a baboon image as mentioned in Section 2.6.2. The location of the occlusions is randomly chosen and unknown to the algorithm.

We use clean images from Subset 1 and Subset 2 (717 images, in normal-to-moderate illumination conditions) as gallery. Images with synthetic occlusions from Subset 3 (453 images, in extreme illumination conditions), Subset 4 (524 images, in more extreme illumination conditions) and Subset 5 (712 images, in the most extreme illumination conditions) are used for testing, respectively. The examples of the images are shown as Figure 2.8a-e. All images are cropped and re-sized to 96×84 pixels.

From the experimental results in [25], the classification accuracies of using different features (i.e., Eigenface [38], Fisherface [39], Laplacianface [158], random projection and downsampled images) are relatively similar within the same SRC framework. However, these features are all holistic features. In our experiments, the results show the performance of SRC can be significantly improved by employing the local features. We first give an example to illustrate how the combination of the SRC with the robust WLD feature helps.

Figure 3.4 is a comparison between the original method SRC-I (left) and the combined method SRC-W (right) using the same probe images (Figure 3.4b and Figure 3.4f) with 40% contiguous occlusions in an extreme illumination condition. The original unoccluded images are shown in Figure 3.4a and Figure 3.4e for comparison. Note that in Figure 3.4c, the left eye is considered as sparse error (bright area) by SRC-I due to the strong shadows. However, only the real occlusions are detected by SRC-W as in Figure 3.4g since the illumination effect is normalised. It is evident that the reconstructed image of SRC-W (Figure 3.4h) is more accurate than that of SRC-I (Figure 3.4d) compared with the corresponding original images (Figure 3.4e and Figure 3.4a). The red entries in Figure 3.4i, j, k and l indicate the index of correct class for the test image (Class 1), respectively. Obviously, the coefficients of SRC-W (Figure 3.4j) are sparser than that of SRC-I (Figure 3.4i) and the coefficients with large magnitude in Figure 3.4j are only associated with the images from the correct class while in Figure 3.4i are not. The smallest residual is correctly associated with Class 1 in Figure 3.4l. In addition, the residual between the test image and the reconstructed images by the correct class is more distinctive in SRC-W than that in SRC-I (Figure 3.4l and Figure 3.4k). The ratio between the two smallest residuals of SRC-W (Figure 3.4l) is 3.3:1 which is much larger than that of 1.1:1 in SRC-I (Figure 3.4k). The similar phenomenon also exists in SSR-I and SSR-W. The combination of SRC based methods with robust local feature is applicable for face recognition in uncontrollable conditions.

Then we evaluate the effectiveness of the proposed SSR model using only the pixel intensity (SSR-I). Table 3.1 shows the comparison of recognition rates between SSR-I and other state-of-the-art approaches on the commonly used testing set Subset 3. CRC-RLS [119] and R-CRC [119] are also reconstruction based methods mentioned in Section 2.4.1. The results of SRC-I, CRC-RLS and R-CRC are cited from the original papers since the experimental settings are the same. SSR-I performs perfectly on the images with up to 30% occlusions with the 100% recognition rate. When the occlusion rises to 40% of the whole image, only 2.2% of the test images are misclassified. Even when half of the image is

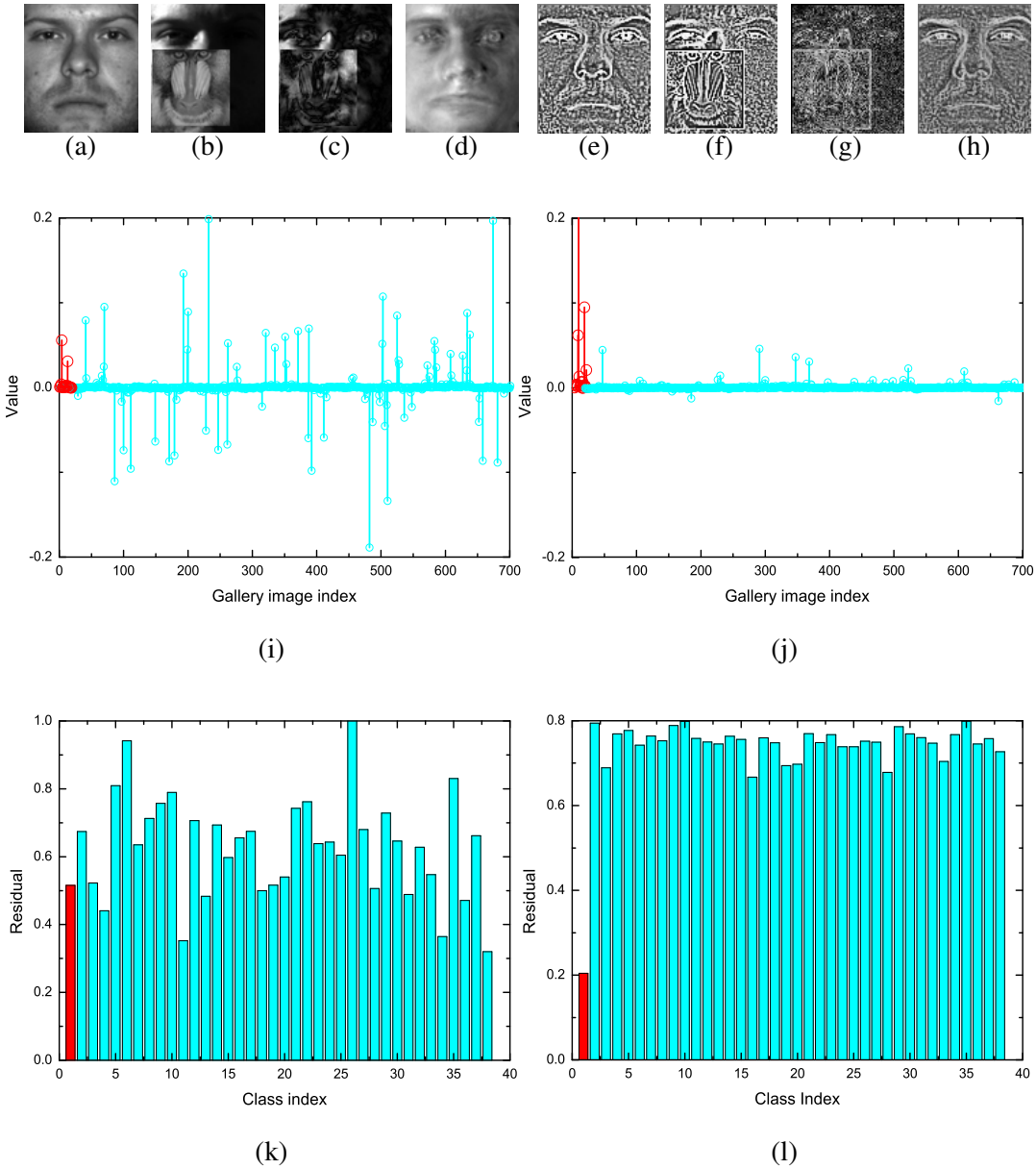


Figure 3.4: The comparison between SRC-I (left) and SRC-W (right). From left to right are: the (a)(e)original images, (b)(f) probe images, (c)(g) estimated sparse errors, (d)(h) reconstructed image, (i)(j) estimated sparse coefficients, and the (k)(l) residuals between the probe image and the reconstructed image recovered by each class. The probe image (b)(f) belongs to Class 1 (indicated in red in (i), (j), (k) and (l)).

occluded, SSR-I still achieves a recognition rate of 85.4%. Note that only the pixel intensity is used in all methods in Table 3.1. It is clear that proposed model, which considers the *structured sparsity*, outperforms all other methods compared. We also test our model using the WLD feature (SSR-W) on Subset 3 and it achieves 100% recognition rate on all levels of occlusions.

Table 3.1: Identification rates (%) on the Subset 3 of the Extended Yale B database

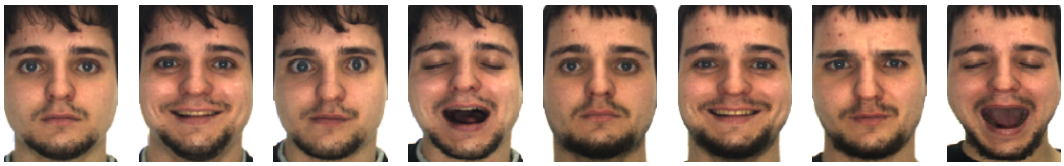
Occlusion	0%	10%	20%	30%	40%	50%
SRC-I [25]	**100**	**100**	99.8	98.5	90.3	65.3
CRC-RLS [119]	**100**	**100**	95.8	85.7	72.8	59.2
R-CRC [119]	**100**	**100**	**100**	97.1	92.3	82.3
Proposed SSR-I	**100**	**100**	**100**	**100**	**97.8**	**85.4**

We next test our proposed methods on the datasets with more extreme conditions: Subset 4 and Subset 5. As shown in Figure 2.8d and e, the images from these two datasets contain significant illumination changes. Very few works test their methods on these two datasets. These images with such extreme illumination effects are difficult to recognise even for human beings, not to mention containing large ratio of synthetic occlusions. We compare our approaches SSR-I and SSR-W with the other three sparse representation based methods as listed at the beginning of this section.

The recognition results are shown in Table 3.2. SSR-I achieves better results on both Subset 4 and Subset 5 compared with the original method SRC-I. But the recognition rates are still very low because of the coupled effect of large illumination changes and occlusions. By using the illumination insensitive WLD feature, our SSR-W achieves consistently better recognition rates on all levels of occlusions while the recognition rates of other methods drop sharply with the increasing level of occlusions. Especially, on Subset 4, SSR-W performs stably, achieving more than 98% recognition rates on all levels of occlusions. Subset 5 contains images with the most extreme illumination changes, SSR-W still achieve an average recognition rate of 95.6% while none of the other methods without using WLD features achieve higher than 50% recognition rate. Even on 50% occlusion with such extreme illumination changes, SSR-W still leads to a recognition rates 88.6%.

Table 3.2: Identification rates (%) on the Subset 4 and Subset 5 of the Extended Yale B database

	Occlusion	0%	10%	20%	30%	40%	50%
Subset 4	SRC-I [25]	86.3	78.5	70.0	53.2	36.7	28.1
	Proposed SSR-I	97.2	93.4	84.8	68.4	53.4	39.9
	SRC-G [113]	95.3	88.8	84.2	76.4	66.5	54.7
	SRC-W	99.4	99.6	**99.4**	99.1	99.1	96.6
	Proposed SSR-W	**99.6**	**99.8**	**99.4**	**99.4**	**99.6**	**98.1**
Subset 5	SRC-I [25]	37.5	26.9	14.3	9.0	7.9	7.3
	Proposed SSR-I	42.6	31.6	23.4	15.3	11.5	10.9
	SRC-G [113]	44.2	31.7	32.0	23.8	21.5	17.5
	SRC-W	98.0	97.5	96.9	**96.9**	91.9	83.0
	Proposed SSR-W	**98.3**	**98.0**	**97.3**	95.8	**95.4**	**88.6**



(a) Unoccluded images

(b) Images occluded by sunglasses and scarves non-uniform illuminations

Figure 3.5: Cropped images from the AR database used in the experiments.

Compared with the recognition rates of SRC-I on both datasets, the average recognition rate of SRC-W increases from 58.8% to 98.9% on Subset 4 and from 17.2% to 94.0% on Subset 5, respectively. This strongly shows that the performance of SRC can be significantly improved by using the local WLD feature when dealing with the coupled illumination changes and occlusion condition.

3.4.2 Face identification with facial disguises and non-uniform illuminations

We next test our approaches on the AR database where the images contain real disguise with non-uniform illumination changes. These images suffer from different variations in facial expressions, illumination conditions and occlusions. Similar to the works in

[16, 17, 25, 115, 124, 151], a subset [159] of the AR database (50 men and 50 women) containing varying illumination conditions, expressions and occlusions is used in our experiments (1,599 images in total, 16 images from each person, except for a corrupted image w-027-14.bmp). For each class, eight unoccluded, frontal view images with various expressions are used as the gallery set. Two separate sets (400 images each) of images simultaneously containing occlusions and left/right side lighting are used for testing. The first set contains images with sunglasses and the other set with scarves (Figure 3.5). Note that this is more challenging than the experiments reported in [25] because each test image includes disguise and non-uniform illumination effect at the same time. All images are cropped and re-sized to 83×60 pixels.

Table 3.3: Identification rates (%) on the AR database

	Sunglasses	Scarves
SRC-I [25]	42.5	29.8
Proposed SSR-I	43.5	31.8
SRC-G [113]	74.8	76.0
SRC-W	85.0	89.5
Proposed SSR-W	**87.5**	**92.0**

Table 3.3 shows the recognition rates. Note that the images with illumination changes are not included in the gallery set. So the illumination conditions in the probe images are quite difference from that in the galley images. When using the pixel intensity, SSR-I performs slightly better than SRC-I. The approaches using WLD (i.e., SRC-W and SSR-W) significantly outperform those using pixel intensity (i.e., SRC-I and SSR-I) and Gabor features (SRC-G). SRC-W achieves an identification rate of 85% on the sunglasses set and 89% on the scarf set, over 40% higher than that of the original method SRC-I [25]. SSR-W achieves the best rate of 87.5% on the sunglasses and 92% on the scarf. The approaches using WLD dramatically outperform the others because of their robustness to the illumination variations which cannot be linearly interpolated with the gallery set.

3.4.3 The effect of cluster size

In this section we investigate the effect of varying cluster sizes on the identification performance. To fairly evaluate the classification accuracy of each class, we assume that all classes have the same number of gallery images. The size of each cluster n_i in the gallery set $\boldsymbol{X}$ in the (3.2), as we analysed before, is the number of gallery images per class. Without loss of generality, we set the structured occlusion dictionary as an equal-size-cluster dictionary where the size of each cluster is d'. We test the proposed SSR-I with different values of $d'(d' = 1, 10, 20, 30, 40, 100)$ on the Extended Yale B database. We use clean images as gallery set and randomly select images with synthetic occlusions in different illumination conditions as the testing set.

The recognition result is shown in Figure 3.6. When $d' = 1$, the structured occlusion dictionary is the same as the identity matrix. From Figure 3.6 we can see that when the size is moderate $(d' = 10, 20, 30)$, using structured occlusion dictionary leads to better identification rates than using the identity matrix $(d' = 1)$, which indicates the effectiveness of the structured occlusion dictionary. When the size is too large $(d' = 100)$ which reduces the flexibility of the occlusion dictionary to represent occlusions with different sizes, the recognition rate decreases. A large cluster size also leads to long running time. As a result, we set d' to 20 in all experiments to strike a good balance between the computational cost and classification accuracy.

3.5 Summary

In this chapter, we have proposed a model considering structured sparsity to simultaneously deal with the coupled condition of large illumination changes and occlusions. Firstly, we propose a structured occlusion dictionary for better modelling contiguous occlusions. Secondly, we employ an illumination insensitive WLD feature for handling severe illumination variations. In our model, we use a minimal number of clusters which only involve the gallery images from the most probable classes to represent a face image, which is

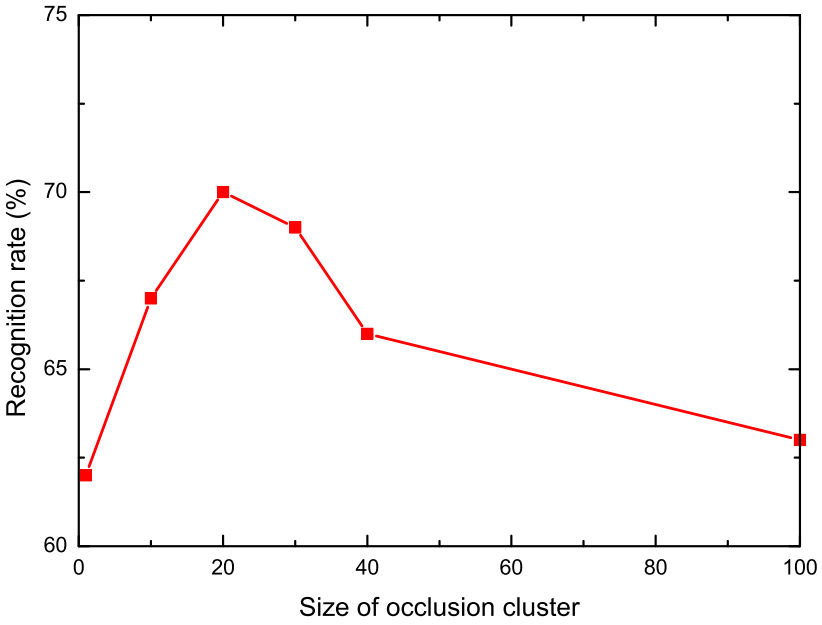


Figure 3.6: Identification rates of SSR-I with different sizes of occlusion cluster

more suitable for classification compared with the original SRC model. The experimental results show that the proposed approach outperforms the state-of-the-art face recognition algorithms for handling multiple influence of illumination changes and occlusions. In addition, our experimental results confirm that employing robust local feature such as WLD in SRC based methods is feasible when handling multiple variations. Our approach provides a baseline for comparison which can help other sophisticated models to verify their performance when dealing with multiple variations.

One limitation for the proposed method as well as the SRC model is that they assume that a large number of gallery images are available for each subject (i.e., the gallery set is an over-completed dictionary). This assumption does not always hold in the real-world application scenarios. So in the following chapters, we will introduce the local patch based methods which work with limited gallery images.

Chapter 4

Dynamic Image-to-Class Warping (DICW)

The reconstruction based approaches introduced in the previous chapters usually require a large number of samples per subject to represent a probe image. However, a sufficient number of samples are not always available in practical scenarios. In this chapter, we deal with occlusions in the other direction as mentioned in Chapter 2 and propose a local matching based method, coined Dynamic Image-to-Class Warping (DICW), for occluded face recognition. In Chapter 2, we mentioned that occlusions give rise to two difficulties for face recognition: 1) the large distance error in the feature space, and 2) the large registration error (i.e., misalignment of images). We will demonstrate that DICW is robust to occlusions and the misalignment.

DICW is motivated by the Dynamic Time Warping (DTW) algorithm [160] which allows elastic matching of two time sequences. In our model, an image is partitioned into patches, which are then concatenated in the raster scan order to form a sequence. Thus, a face is represented by a patch sequence which contains the *order information* of facial features. DICW calculates the *Image-to-Class* distance between a query face and those of an enrolled subject by finding the optimal alignment between the query sequence and all enrolled sequences of that subject. It allows elastic matching in both *time* and *with-class*

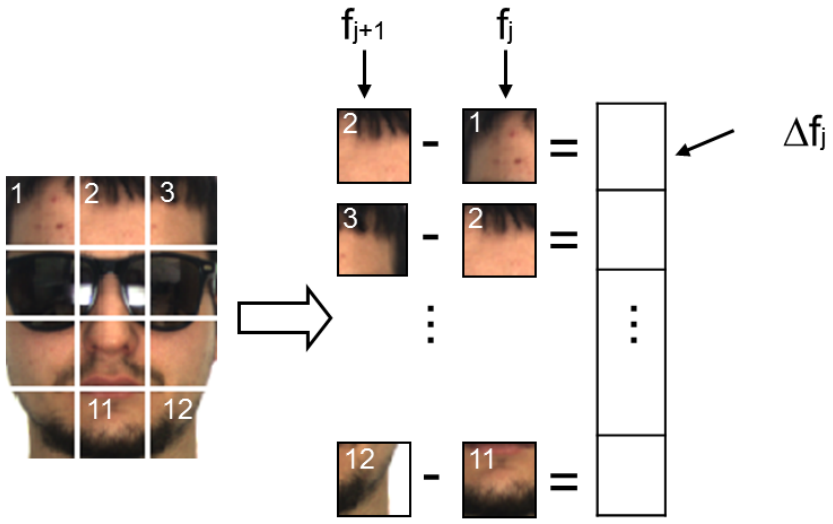


Figure 4.1: The image representation of DICW.

dimensions.

In this chapter, the proposed DICW, from image representation, modelling to implementation, is described in Section 4.1 to 4.3. Extensive experiments are demonstrated in Section 4.4. The discussions of the DICW's robustness to misalignment and the settings of parameters are presented in Section 4.5. Further analysis about why the proposed method works; when and why it will fail and how to improve it are discussed in Section 4.6. Finally, the work is concluded in Section 4.7. Note that symbols used are only valid within this chapter.

4.1 Image representation

In this work, an image is partitioned into J non-overlapping patches of $d \times d'$ pixels. Those patches are then concatenated in the raster scan order (i.e., from left to right and top to bottom) to form a single sequence. The reason for doing so is that the forehead, eyes, nose, mouth and chin are located in the face in a natural order, which does not change despite the occlusions or imprecise registrations (i.e., small rotations). This *spatial facial order*, which is contained in the patch sequence, can be seen as the *temporal order* in a time sequence.

In this way, a face image can be seen as a time sequence so the image matching problem can be handled by time series analysis techniques like DTW [160]. Throughout the work we will use the terms *temporal order* and *spatial order* interchangeably.

Let f_j be the vector of the intensity of the pixels in the j-patch. A difference patch $\triangle f_j$ is computed (Figure 4.1) by subtracting f_j from its immediate neighbouring patch f_{j+1} as:

$$\triangle f_j = f_{j+1} - f_j \tag{4.1}$$

where $j \in \{1, 2, ..., J-1\}$. Note that here the length of the *difference patch* sequence is $J-1$.

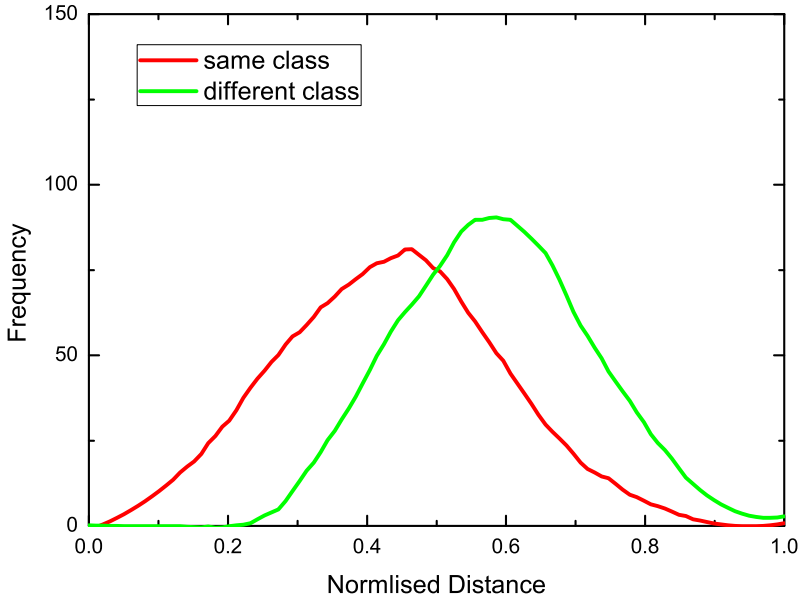
A difference patch $\triangle f_j$ actually can be viewed as the approximation of the first-order derivative of adjacent patch f_{j+1} and f_j. The salient facial features which represent detailed textured regions such as eyes, nose and mouth can be enhanced since the first-order derivative operator is sensitive to edges.

We use 3,200 occluded-unoccluded image pairs of the same class and different classes from the AR database [132], respectively (6,400 pairs in total) to calculate the image distance distributions[1]. As shown in Figure 4.2, the distance distributions of the same and different classes are separated more widely when using the difference patches (Figure 4.2b).

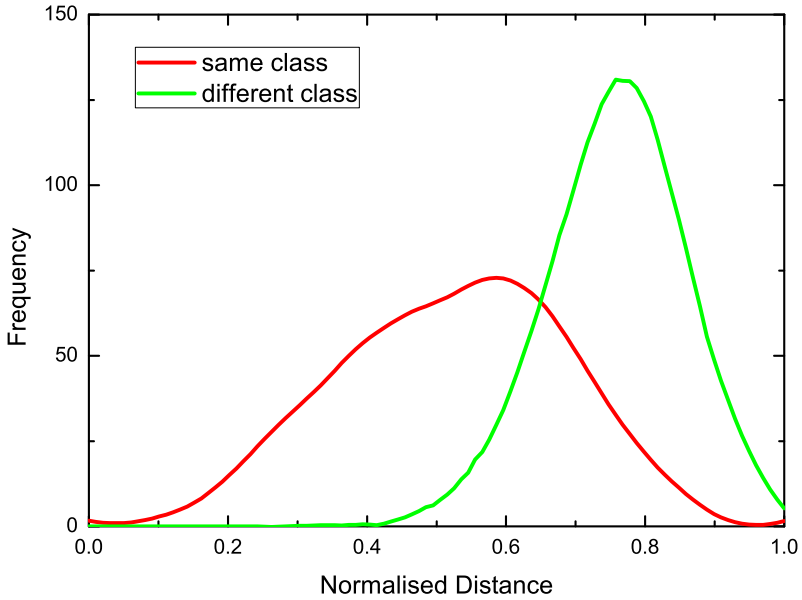
4.2 Modelling

Face matching is implemented by defining a distance measurement between sequences and using the distance as the basis for classification. Generally, a small distance is expected if two sequences are similar to each other. DICW is motivated by the DTW algorithm [160] which allowing elastic matching of two time sequences. It has been successfully applied to the area of speech recognition [160]. Here an example is used to quickly illustrate the main idea of DTW (more details of the algorithm can be found in [160]). As shown in Figure

[1]We use Euclidean distance as measurement. The image size is 83×60 pixels and the patch size is 5×5 pixels



(a)



(b)

Figure 4.2: Distributions of face image distance of the same and different classes. Using the difference patch (b), the distance distribution of the same class and that of the different classes are separated more widely compared with those using the original patch (a).

4.3, there are two sequences (each number indicates a data point):

$$\begin{aligned} A =& (a_1, a_2, a_3, a_4, a_5) = (3, 1, 10, 3, 2) \\ B =& (b_1, b_2, b_3, b_4, b_5) = (3, 1, 2, 10, 3). \end{aligned} \tag{4.2}$$

The Euclidean distance (i.e., using point-wise matching, Figure 4.3a) between them is

$$\sqrt{(a_1 - b_1)^2 + ... + (a_5 - b_5)^2} = \sqrt{0 + 0 + 64 + 49 + 1} \approx 10.68$$

which is a bit large for these two similar sequences. However, if we *warp* these two sequences in a non-linear way by shrinking or expanding them along the time axis during matching (i.e., allows flexible correspondences), the distance between A and B can be largely reduced [2] to 2 (Figure 4.3b). DTW, which is based on this idea, calculates the distance between two time sequences by finding the optimal alignment between them with the minimal overall cost. This will help to reduce the distance error caused by some *noise* data points and ensure that the distance between similar sequences is relatively small. In addition, the *temporal order* is considered during matching, thus cross-matching (which reverses the order of data points) is not allowed even though it can lead to shorter distance (Figure 4.3c). Especially for face recognition, this is reasonable since the order of facial features should not be turned back.

Adopting this idea for face recognition, we want to find the optimal alignment between face sequences while minimising the distance caused by occluded patches. In this work, instead of finding alignment between two sequences, we seek alignment between a sequence and the sequence *set* of a given class (i.e., subject). A probe image consisting of M patch features is denoted by $\boldsymbol{P} = (\boldsymbol{p}_1, ..., \boldsymbol{p}_m, ..., \boldsymbol{p}_M)$. Here $\boldsymbol{P}$ is an ordered list where each element $\boldsymbol{p}_m$ is a patch feature vector (e.g., $\triangle f$ in Section 4.1). The gallery set of a given class containing K images is denoted by $\boldsymbol{G} = \{\boldsymbol{G}_1, ..., \boldsymbol{G}_k, ..., \boldsymbol{G}_K\}$. The k-th gallery image is similarly represented as a sequence of N patch features as

[2] Computation details see [160]

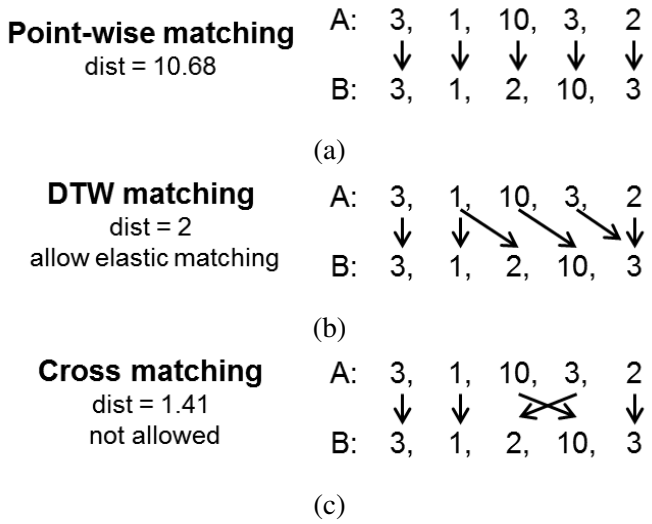


Figure 4.3: Various ways of sequence matching. a) Point-wise matching, b) DTW matching, and c) cross matching.

$\boldsymbol{G}_k = (\boldsymbol{g}_{1k}, ..., \boldsymbol{g}_{nk}, ..., \boldsymbol{g}_{Nk})$ where $\boldsymbol{g}_{nk}$ represents a patch feature vector like $\boldsymbol{p}_m$. Note that the number of patches in the probe image and that in the gallery image can be different (i.e., the values of M and N can be different) since the DTW model [160] is able to deal with sequences with different lengths.

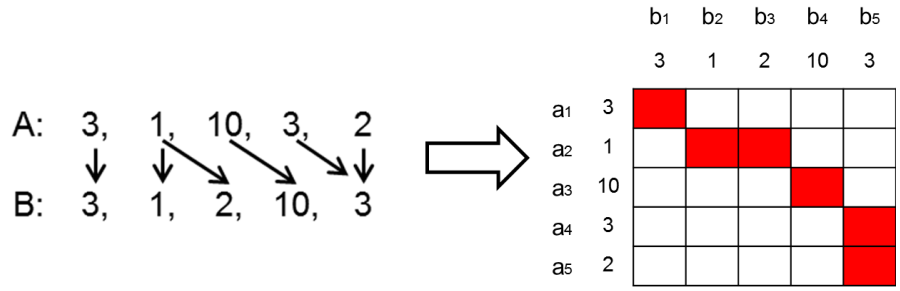
A warping path $\boldsymbol{W}$ indicating the matching correspondence of patches between $\boldsymbol{P}$ and $\boldsymbol{G}$ with T warping steps in time axis is defined as $\boldsymbol{W} = (w(1), ..., w(t), ..., w(T))$ with:

$$w(t) = (m, n, k) : \{1, 2, ..., T\} \rightarrow \{1, 2, ..., M\} \times \{1, 2, ..., N\} \times \{1, 2, ..., K\} \quad (4.3)$$

where $\times$ indicates the Cartesian product operator and $\max\{M, N\} \leqslant T \leqslant M + N - 1$. $w(t) = (m, n, k)$ is a triple indicating that patch $\boldsymbol{p}_m$ is matched to patch $\boldsymbol{g}_{nk}$ at step t.

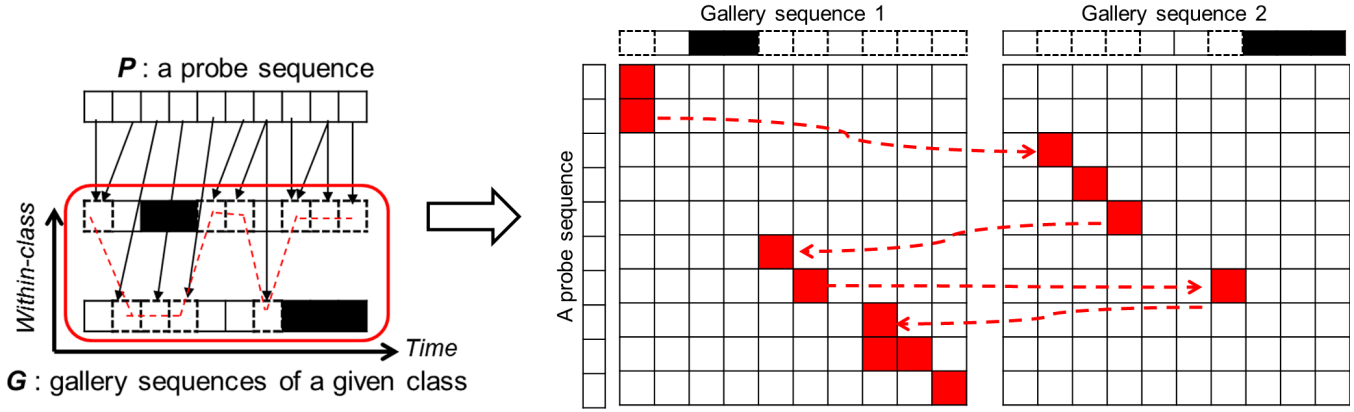
Similar to the DTW model [160], $\boldsymbol{W}$ in DICW satisfies the following four constraints:

1. Boundary: $w(1) = (1, 1, k)$ and $w(T) = (M, N, k')$. The path starts at matching $\boldsymbol{p}_1$ to $\boldsymbol{g}_{1k}$ and ends at matching $\boldsymbol{p}_M$ to $\boldsymbol{g}_{Nk'}$. Note that no restrictions are placed on k and k'. From step 1 to T, k and k' can be any value from 1 to K since the probe patch can be



$$\boldsymbol{W} = ((1,1),(2,2),(2,3),(3,4),(4,5),(5,5)), T = 6$$

(a)



$$\boldsymbol{W} = ((1,1,1),(2,1,1),(3,2,2),(4,3,2),(5,4,2),(6,5,1),(7,6,1),(7,7,2),(8,8,1),(9,8,1),(9,9,1),(10,10,1)), T = 12$$

(b)

Figure 4.4: An illustration of warping path in a) DTW and the b) proposed DICW. The arrows indicate the matching correspondence. The dashed line marks the optimal warping path. The black blocks indicate the occluded patches.

matched with patches from all K gallery images.

2. Monotonicity: Given $w(t) = (m, n, k)$, the preceding triple $w(t-1) = (m', n', k')$ satisfies that $m' \leqslant m$ and $n' \leqslant n$. The warping path preserves the *temporal order* and increase monotonically.

3. Continuity: Given $w(t) = (m, n, k)$, the preceding triple $w(t-1) = (m', n', k')$ satisfies that $m - m' \leqslant 1$ and $n - n' \leqslant 1$. The indexes of the path increase by 1 in each step, which means that each step makes smooth transitions along the *time* dimension.

4. Window constraint: Given $w(t) = (m, n, k)$, it satisfies $|m - n| \leqslant l$ where $l \in \mathbb{N}^+$ is the window width [160]. The window constraint is designed to reduce the computational cost of DICW. But it is also meaningful for the specific face recognition problem since a probe patch (e.g., eye) should not match to a patch (e.g., mouth) too far away. The window with a width l is able to constrain the warping path within an appropriate range.

These constraints are extended from the constraints of the DTW algorithm. However, they are also meaningful in the context of face recognition with the image representation defined in Section 4.1. Our method represents a face image as a patch sequence thus the image matching problem can be solved by the time series analysis technique.

In order to explain the concept of *warping path*, we take the aforementioned sequences A and B as an example. In Figure 4.4a, each grid on the right hand side indicates a possible matching correspondence. The indexes of the red grids indicate the matching between A and B by DTW (i.e., the optimal warping path with the minimal matching cost) as shown in the left part (here $T = 6$). Likewise, the same procedure of DICW is shown in Figure 4.4b. Compared with DTW, an additional index is added in the warping step of DICW to index different gallery sequences. In this way, the *warping* is performed in two directions: 1) a probe sequence $\boldsymbol{P}$ is aligned to a set of gallery sequences $\boldsymbol{G}$ according to the *time* dimension (maintaining the *facial order*) and 2) simultaneously, at each warping step, each patch in $\boldsymbol{P}$ can be matched with any patch among all gallery sequences along

the *within-class* dimension. Our method allows elastic match in both of the aforementioned two directions.

We define the local distance [160] $C_{m,n,k} = d(\boldsymbol{p}_m, \boldsymbol{g}_{nk})$ as the distance between two patches $\boldsymbol{p}_m$ and $\boldsymbol{g}_{nk}$. $d(\cdot)$ can be any distance metric such as the Euclidean distance or the Cosine distance. The overall matching cost of $\boldsymbol{W}$ is the sum of the local distance of each warping step:

$$S(\boldsymbol{W}) = \sum_{t=1}^{T} C_{w_t} \tag{4.4}$$

The optimal warping path $\boldsymbol{W}^*$ (i.e., the red grid path in Figure 4.4b) is the path that minimises $S(\boldsymbol{W})$. The *Image-to-Class* distance between $\boldsymbol{P}$ and $\boldsymbol{G}$ is simply the overall cost of $\boldsymbol{W}^*$:

$$dist_{DICW}(\boldsymbol{P}, \boldsymbol{G}) = \min_{\boldsymbol{W}} \sum_{t=1}^{T} C_{w_t} \tag{4.5}$$

After computing $dist_{DICW}$ between $\boldsymbol{P}$ and each enrolled subject in the database, a classifier such as the Nearest Neighbour classifier can be adopted for classification based on $dist_{DICW}$.

4.3 Implementation through Dynamic Programming

To compute $dist_{DICW}(\boldsymbol{P}, \boldsymbol{G})$ in (4.5), one could test every possible warping path but with a high computational cost. (4.5) can be solved efficiently using *Dynamic Programming*. A three-dimensional matrix $\boldsymbol{D} \in \mathbb{R}^{M \times N \times K}$ is created to store the cumulative distance. The element $D_{m,n,k}$ stores the cost of the optimal warping path of matching the first m probe patches to the set of first n patches of each gallery sequence and at the same time the m-th patch $\boldsymbol{p}_m$ is matched to the patch from the k-th gallery image. The calculation of the final optimal cost $dist_{DICW}(\boldsymbol{P}, \boldsymbol{G})$ is based on the results of a series of predecessors. $\boldsymbol{D}$ can be

computed recursively as:

$$D_{m,n,k} = \min \left\{ \begin{array}{l} \boldsymbol{D}_{\{(m-1,n-1)\}\times\{1,2,\ldots,K\}}, \\ \boldsymbol{D}_{\{(m-1,n)\}\times\{1,2,\ldots,K\}}, \\ \boldsymbol{D}_{\{(m,n-1)\}\times\{1,2,\ldots,K\}} \end{array} \right\} + C_{m,n,k} \tag{4.6}$$

where the initialisation is done by extending $\boldsymbol{D}$ as an $(M+1)\times(N+1)\times K$ matrix and setting $\boldsymbol{D}_{0,0,\cdot} = 0, \boldsymbol{D}_{0,n,\cdot} = \boldsymbol{D}_{m,0,\cdot} = \infty$. Thus, $dist_{DICW}(\boldsymbol{P}, \boldsymbol{G})$ can be obtained as follows:

$$dist_{DICW}(\boldsymbol{P}, \boldsymbol{G}) = \min_{k\in\{1,2,\ldots,K\}} \{\boldsymbol{D}_{M,N,k}\} \tag{4.7}$$

Different from the point-wise matching (in which each patch is viewed as a data point), our method tries every possible warping path under the temporal constraints and selects the one with the minimal overall cost. So the warping path with a large distance error will not be selected. The *Image-to-Class* distance is the *globally* optimal cost for matching. Although the occlusions are not directly removed, avoiding large distance error by warping is helpful for classification, as our experimental results show (see Section 4.4).

In addition, a patch of the probe image can be matched to the patches of K different gallery images of the same subject/class. Because the chance that all patches at the same location of the K images are occluded is low, the chance that a probe patch is compared to an unoccluded patch at the same location is thus higher. When occlusions occur in probe and/or gallery images, the *Image-to-Image* distance may be large. However, our model is able to exploit the information from different gallery images and reduce the effect of occlusions (Figure 4.5). Algorithm 2 summarises the procedure of computing the *Image-to-Class* distance between a probe image and a class. l is the window width and usually set to 10% of $\max\{M, N\}$ [160]. Computational complexity is analysed in Section 4.5.6.

Once the cumulative distance matrix $\boldsymbol{D}$ is computed, the warping path $\boldsymbol{W}$ can be constructed by backtracking from the end element $\boldsymbol{D}_{M,N,\cdot}$ to the start element $\boldsymbol{D}_{1,1,\cdot}$ by using the greedy strategy. Algorithm 3 describes the procedure of finding the warping path.

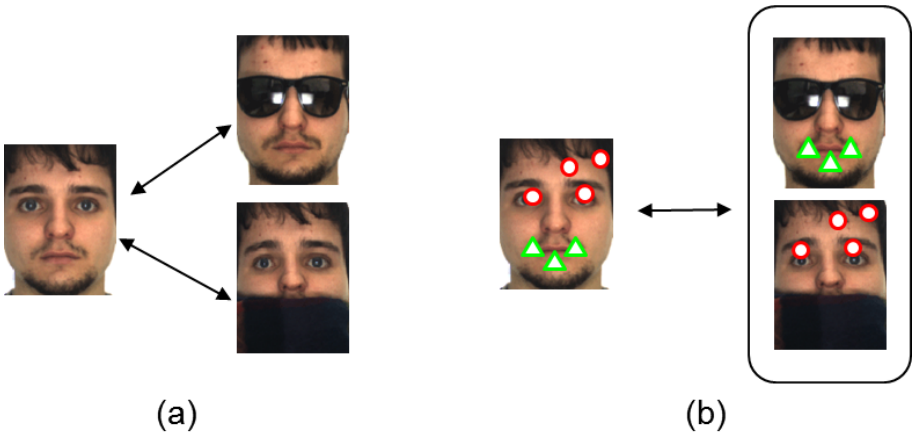


Figure 4.5: The illustration of a) the *Image-to-Image* and b) the *Image-to-Class* matching. Matched features are indicated by the same symbol.

Algorithm 2 Dynamic Image-to-Class Warping distance DICW($\boldsymbol{P}, \boldsymbol{G}, l$)

Input:

$\boldsymbol{P}$: a probe sequence with M patches;
$\boldsymbol{G}$: a set of K gallery sequences (each with N patches) of a given class;
l: the window width;

Output:

$dist_{DICW}$: the *Image-to-Class* distance between $\boldsymbol{P}$ and $\boldsymbol{G}$;

Set each element in $\boldsymbol{D}$ to ∞;
$\boldsymbol{D}[0, 0, 1:K] = 0$;
$l = \max\{l, |M - N|\}$;
Compute the local distance matrix $\boldsymbol{C}$;
for $m = 1$ to M **do**
  for $n = \max\{1, m - l\}$ to $\min\{N, m + l\}$ **do**
    $minNeighbour = \min \left\{ \begin{array}{l} \boldsymbol{D}[m-1, n-1, 1:K], \\ \boldsymbol{D}[m-1, n, 1:K], \\ \boldsymbol{D}[m, n-1, 1:K] \end{array} \right\}$;
    for $k = 1$ to K **do**
      $D[m, n, k] = minNeighbour + C[m, n, k]$;
    end for
  end for
end for
$dist_{DICW} = \min\{\boldsymbol{D}[M, N, 1:K]\}$;
return $dist_{DICW}$;

In the classification, only the DICW distance $dist_{DICW}$ from Algorithm 2 is used. But the constructed warping path by Algorithm 3 can be used for other analysis.

Algorithm 3 Warping path construction($\boldsymbol{D}$)

Input:

$\boldsymbol{D}$: a $(M+1) \times (N+1) \times K$ cumulative distance matrix;

Output:

$\boldsymbol{W}$: the warping path;

$m = M; n = N; k = 0; \boldsymbol{W} \leftarrow$ array;
while $m > 1 \&\& n > 1$ **do**
  if $m == 1$ **then**
    $n = n - 1$;
  else if $n == 1$ **then**
    $m = m - 1$;
  else
    $[value_1, index_1] = \min\{\boldsymbol{D}[m-1, n, 1:K]\}$;
    $[value_2, index_2] = \min\{\boldsymbol{D}[m, n-1, 1:K]\}$;
    $[value_3, index_3] = \min\{\boldsymbol{D}[m-1, n-1, 1:K]\}$;
    $[minValue, minIndex] = \min\{value_1, value_2, value_3\}$;
    if $minIndex == 1$ **then**
      $m = m - 1; k = index_1$;
    end if
    if $minIndex == 2$ **then**
      $n = n - 1; k = index_2$;
    end if
    if $minIndex == 3$ **then**
      $m = m - 1; n = n - 1; k = index_3$;
    end if
  end if
  $\boldsymbol{W} \leftarrow (m, n, k)$;
end while
return $\boldsymbol{W}$;

4.4 Experimental analysis

In this section, we evaluate the proposed method on four databases (FRGC [52], AR [132], TFWM [135] and LFW [134]). We perform identification tasks according to the three cases (i.e., **Uvs.O**, **Ovs.U** and **Ovs.O**) described in Chapter 2. We first consider the scenario where occlusions occuring only in probe images (i.e., **Uvs.O**) and test our method

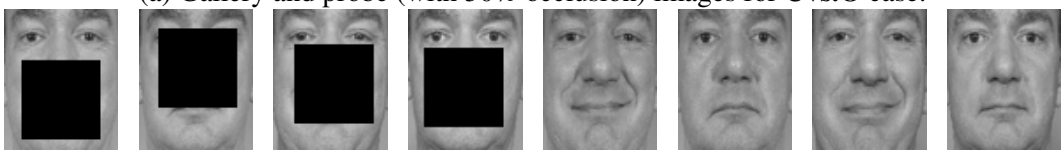
using different number of gallery images per subject. Next, we consider the situation that occlusions exist in the gallery images, which is a case most of the current works do not take account. We fix the number of gallery images per subject and conduct experiments step by step: firstly the probe images are not occluded (i.e., **Ovs.U**); and then both the gallery and probe images are occluded (i.e., **Ovs.O**). Note that, for comparison purposes, the experiments on the FRGC and the AR databases also include the case that no occlusion is present in both gallery and probe images to confirm that our DICW is also effective in general conditions. In addition, we also extend DICW to verification tasks with faces containing large variations. The effect of patch size, the discriminative power of the difference patch, the robustness to misalignment and the computational complexity are also discussed.

Note that in all experiments, the gallery image set is disjoint with all probe sets. Considering that the gallery and probe images are at the same scale, in the experiments, the probe images and the gallery images are partitioned into the same number of patches, i.e., $M = N$ as defined in Section 4.2. As recommended in the work reported in [161], the Euclidean distance and the Cosine distance are used as local distance metrics for the pixel intensity feature and the LBP feature [50], respectively.

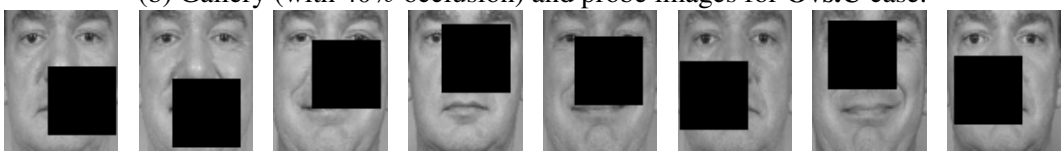
We quantitatively compare DICW with some representative methods in the literature: the supervised linear SVM [162] using PCA [38] for feature extraction (PCA+SVM), the reconstruction based SRC [25] as introduced in Chapter 3, the *Image-to-Class* distance based Naive Bayes Nearest Neighbour (NBNN) [163] as ours, and the baseline, Hidden Markov models (HMM) [42] which also considers the order information in a face. We use the difference patch representation as defined in Section 4.1 in NBNN and our DICW. For comparison purpose, we also report the results of using the original patche (referred to OP-NBNN and OP-Warp, respectively).

Note that NBNN is a local patch based method which also exploits the *Image-to-Class* distance. But it does not consider the spatial relationship between patches like ours. To improve the performance, a location weight α [163] is used in NBNN to constrain

(a) Gallery and probe (with 50% occlusion) images for **Uvs.O** case.



(b) Gallery (with 40% occlusion) and probe images for **Ovs.U** case.



(c) Gallery and probe images with 30% occlusion for **Ovs.O** case.

Figure 4.6: Sample images from the FRGC database with randomly located occlusions used in the experiments. In each row, the first four images are used as gallery and the remaining four images are used as probe. Note that images in the gallery set are different from those in the probe sets and the occlusion locations are different in gallery and probe images in the **Ovs.O** case.

matching patches according to their locations. We tested different values of α and found that the performance of NBNN is highly dependent on the value of α and different testing data (e.g., different occlusion level, location) requires different value even within the same database. So we also reported the best result for each test with the empirically best values of α (as OP-NBNN-ub and NBNN-ub). The performance of OP-NBNN-ub and NBNN-ub can be seen as the upper bound of the performance of NBNN, which is a competitive comparison for our DICW.

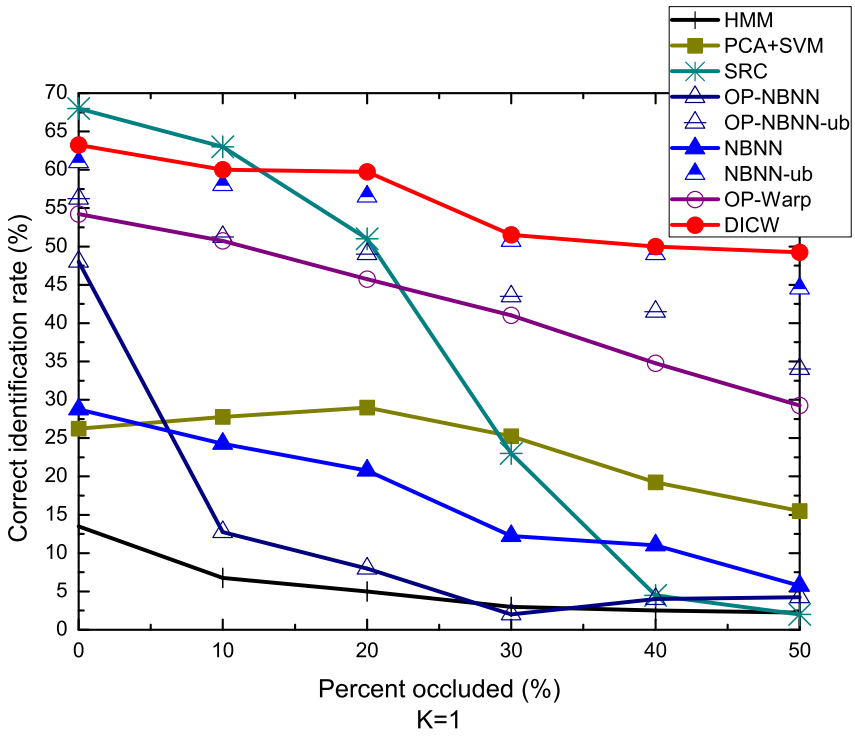
4.4.1 Face identification with randomly located occlusions

We first evaluate the proposed method on the FRGC database [52] (version 2.0, Experiment 4) with randomly located occlusions. Note that in each image, the location of the occlusion is randomly chosen and unknown to the algorithm. Especially, in the **Ovs.O** scenario, the locations of occlusions in the gallery images are different from those in the probe images (Figure 4.6c). We use these images with randomly located occlusions to evaluate

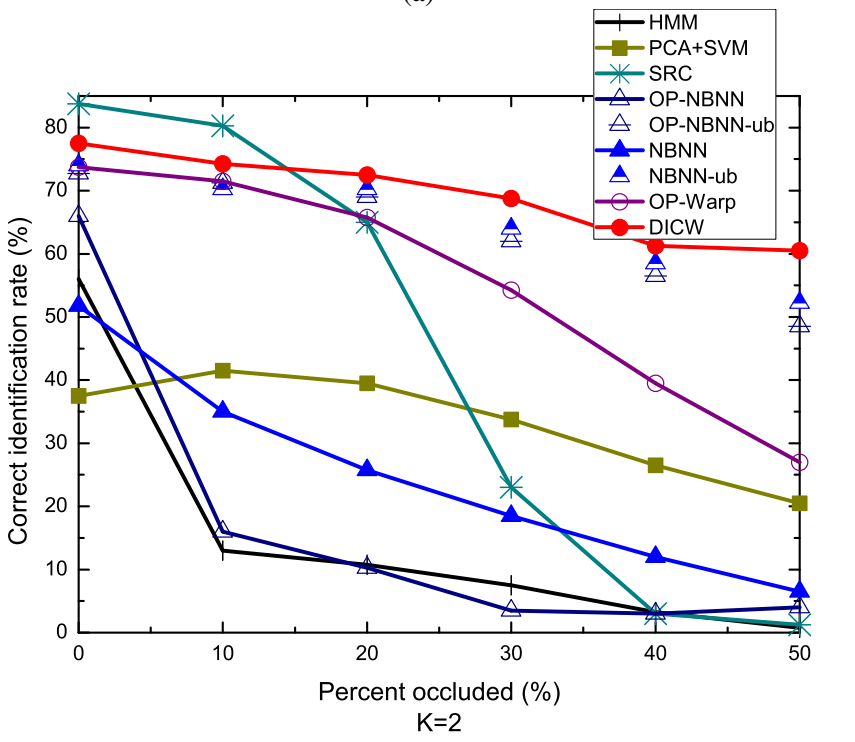
the effectiveness of our method when there is no prior knowledge of the occluded location.

Similar to the work in [17], an image set of 100 subjects (eight images in two sessions are selected for each subject), is used in our experiments. Similar to Chapter 3, we simulate the randomly located occlusions for the probe images. We create an occluded image set by replacing a randomly located square patch (size from 10% to 50% of the original image) from each image in the original image set with a black block. We design the experiments according to the three occlusion scenarios: **Uvs.O**, **Ovs.U** and **Ovs.O** (Figure 4.6). There are 2,400 test samples for each scenario. All images are cropped and re-sized to 80×65 pixels and the patch size is 6×5 pixels (the effect of patch size is discussed in Section 4.5.1).

Uvs.O: For each subject, we select $K = 1, 2, 3$ and 4 unoccluded images respectively to form the gallery sets and use the other four images with synthetic occlusions as the probe set. Fig. 4.7 shows the recognition results with different values of K. The correct identification rates of all methods increase when more gallery images are available (i.e., greater value of K). When there are multiple gallery images per class and no occlusion (level=0%) in images, HMM performs better than the supervised method SVM and the local matching based NBNN. But its performance is significantly affected by the increasing occlusions. In addition, when $K = 1$, HMM performs worst among all methods since there are not enough gallery images to train a HMM for each class. For NBNN and DICW, using the difference patch achieves better results than using the original patch (i.e., OP-NBNN and OP-Warp). Especially, by comparing DICW with OP-Warp, and NBNN with OP-NBNN, it can be found that difference patches improve the results of DICW more significantly than that of NBNN. As introduced in Section 4.1, the difference patches are generated by the spatially continuous patches so they enhance the *order information* within a patch sequence, which is compatible with our DICW. With the empirically best values of location weights, NBNN-ub and OP-NBNN-ub perform better than SVM. When $K = 1, 2, 3$ and 4, the average rates for the six occlusion levels of DICW are 2.3%, 4.3%, 5.5% and 4.4% better than that of NBNN-ub, respectively. When the occlusion level $= 0\%$, the performance

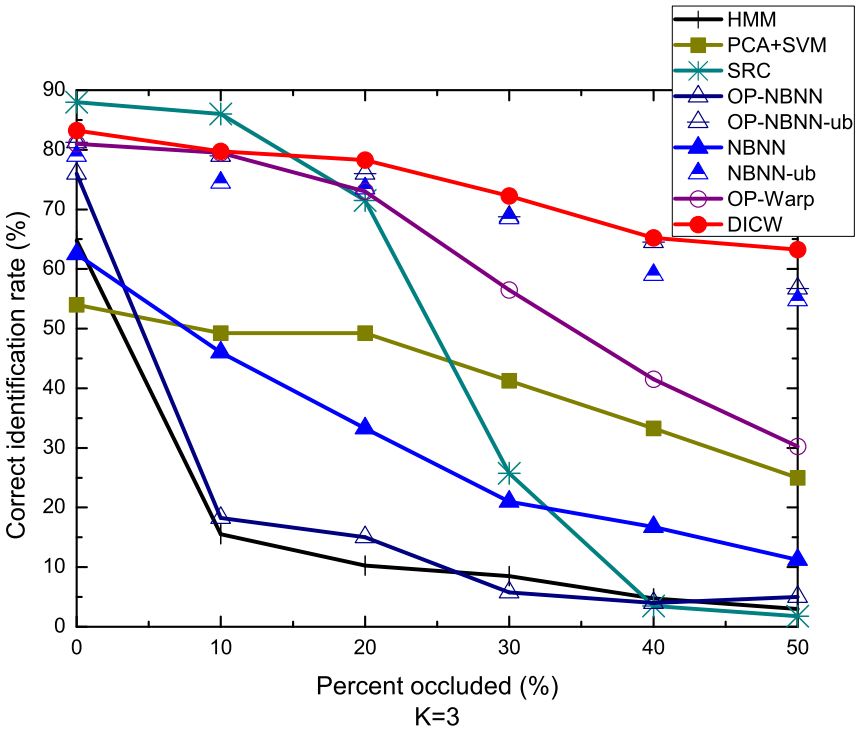


(a)

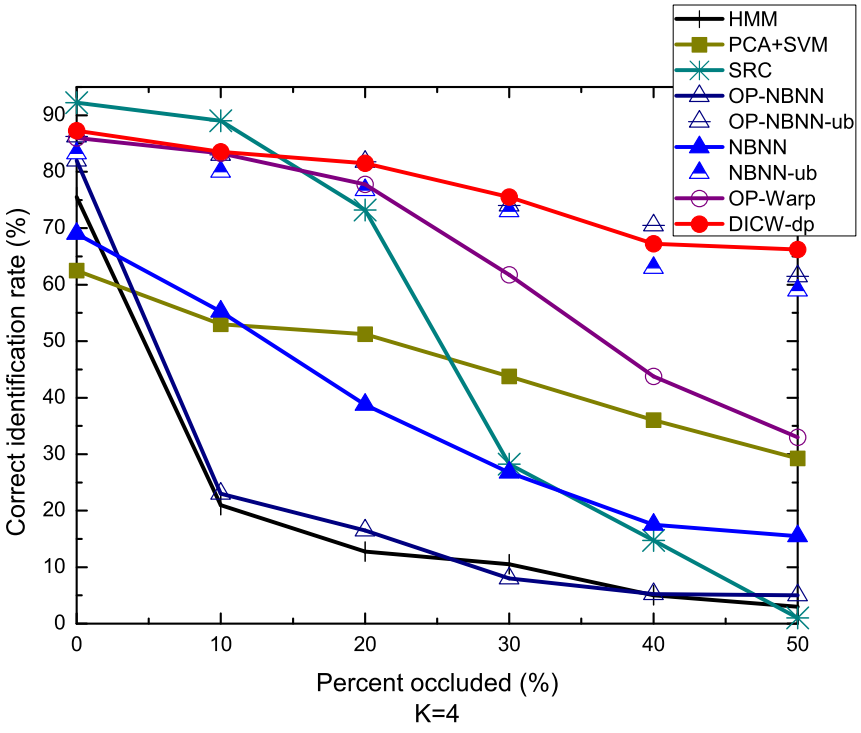


(b)

Figure 4.7: **Uvs.O**: identification results on the FRGC database with different number of gallery images (K) per subject.



(c)



(d)

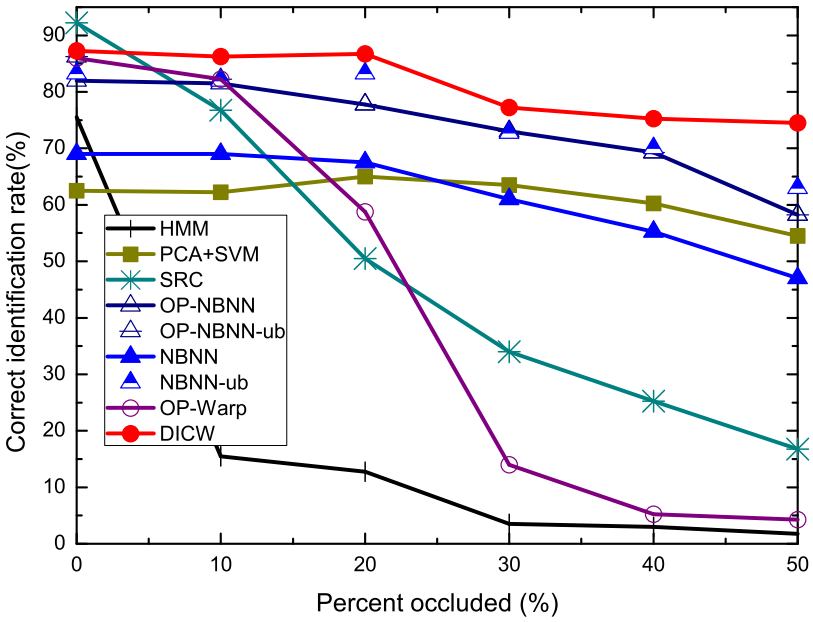
Figure 4.7: **Uvs.O**: identification results on the FRGC database with different number of gallery images (K) per subject (con't).

of SRC is better than DICW. However, the performance drops sharply when the degree of occlusion increases, because the number of gallery images per subject is limited for reconstruction. When $K = 1$, the *Image-to-Class* distance degenerates to the *Image-to-Image* distance. DICW, which allows *time warping* during matching, still achieves better results while the level of occlusion increases.

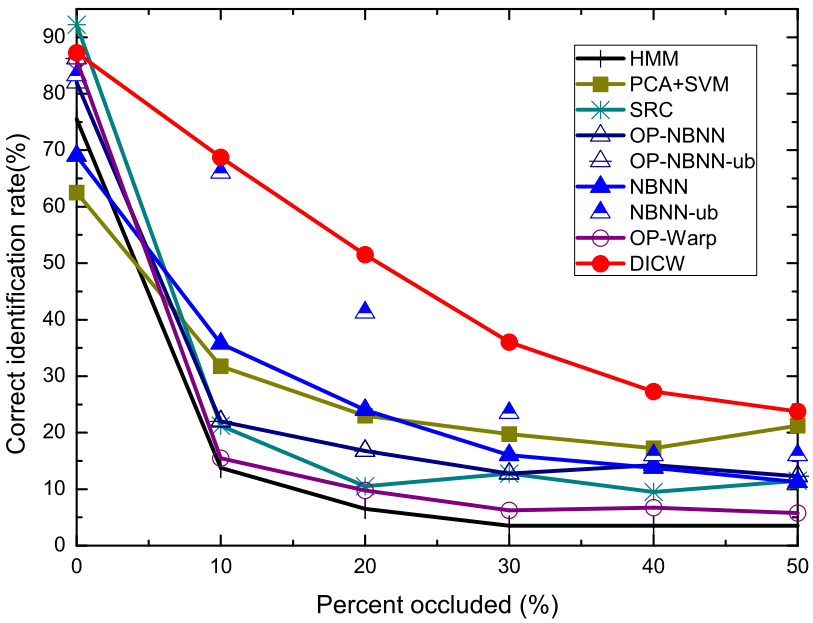
Ovs.U and **Ovs.O**: We fix the value of K to 4 and consider that occlusions exist in the galley set. For each occlusion level (from 0% to 50%), we conduct experiments with the following settings: 1) 400 occluded images (four images per subject) from the original set as the gallery set and 400 images from the unoccluded set as the probe set (**Ovs.U**) and 2) 400 occluded images as the gallery set and 400 occluded images as the probe set (**Ovs.O**). Note that the images in the gallery set are different from those in the probe sets. Fig. 4.8 shows the recognition results. The methods (e.g., HMM, SVM, SRC) which include occluded gallery images for training/modelling perform poorly in these two cases. NBNN does not perform consistently in **Ovs.U** and **Ovs.O**. Using the original patch (i.e., OP-NBNN) performs better than using the difference patch (i.e., NBNN) in **Ovs.U**. For DICW, using the difference patch is always better than using the original patch (i.e., OP-Warp). This confirms that the difference patch works better with DICW, as analysed before. DICW outperforms the best of NBNN (i.e., NBNN-ub) by a larger margin of 5.5% (Fig. 4.8a) and 8.1% (Fig. 4.8b) on average than that (4.4% in Fig. 4.7d) in the **Uvs.O** tested with $K = 4$. These results confirm the effectiveness and robustness of DICW when the gallery and probe images are occluded. On the whole, our method performs consistently and outperforms other methods in all three occlusion cases.

4.4.2 Face identification with facial disguises

We next test the proposed method on the AR database [132] as introduced in Chapter 2. First, we consider that no occlusion is present in both gallery and probe sets. Next, we conduct experiments according to the three occlusion cases (i.e., **Uvs.O**, **Ovs.U** and **Ovs.O**). DICW does not rely on the prior knowledge of the occlusion. We will demonstrate that it

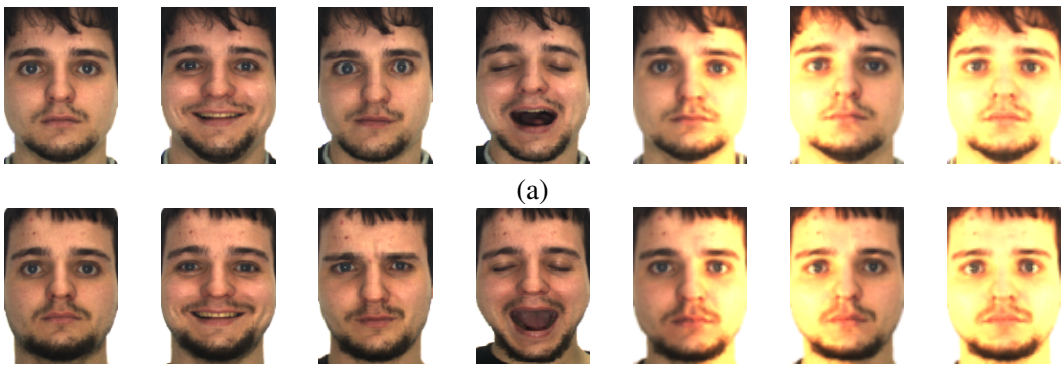


(a)



(b)

Figure 4.8: a) **Ovs.U** and b) **Ovs.O**: identification rates on the FRGC database with occlusions in gallery or/and probe sets.



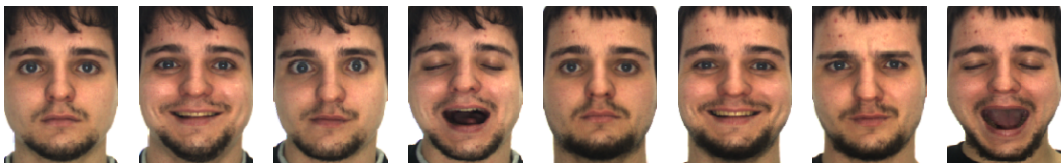
(a)

(b)

Figure 4.9: Cropped images from the AR database *without occlusion*: a) gallery and b) probe.

works well in both general and difficult situations later. All images are cropped and re-sized to 83×60 pixels and the patch size is 5×5 pixels.

Without occlusion: We have evaluated the performance of DICW when no occlusion exists in both gallery and probe sets in Section 4.4.1 (i.e., occlusion level = 0% in the experiments). In this section, we adopt the setting in [25] using images without occlusions to further test DICW. For each subject, 14 images are chosen (four neutral faces with different illumination conditions and three faces with different expressions in each session). Seven images from Session 1 are used as the gallery set and the other seven from Session 2 as the probe set (Figure 4.9). Table 4.1 shows the identification rates. HMM does not perform as well as others. This may be due to other variations such as illumination and expression changes in the training images. Again, the difference patch does not improve NBNN comparing with the original patch (i.e., OP-NBNN). With the empirically best values of location weights, the difference patch (i.e., NBNN-ub) is 3.7% better than the original patch (i.e., OP-NBNN-ub). For DICW, using the difference patch is 3.1% better than using the original patch (OP-Warp). As analysed in Section 4.4.1, the difference patch can enhance the relative *order* of adjacent patches, the results in Table 4.1 also indicates that the difference patch is more compatible with these methods which considers the *order information*. When there is no occlusion in the gallery and probe images, both



(a) Unoccluded images with different expressions (i.e., neutral, smile, anger and scream)

(b) Occluded images with sunglasses and scarves.

Figure 4.10: Sample images from the AR database for the occlusion test: **Uvs.O**, **Ovs.U** and **Ovs.O**.

reconstruction based method (e.g., SRC) and local matching based methods (e.g., NBNN and DICW) achieve relatively satisfactory results. DICW significantly outperforms NBNN and is still slightly better than the upper bound of NBNN (i.e., NBNN-ub).

Table 4.1: Identification results on the AR database without occlusion (K=7)

Method	Correct identification rate (%)
HMM [42]	66.5
PCA+SVM [162]	89.7
SRC [25]	92.0
OP-NBNN [163]	89.6
OP-NBNN-ub [163]	92.0
NBNN [163]	85.3
NBNN-ub [163]	95.7
OP-Warp	93.6
Proposed DICW	**96.7**

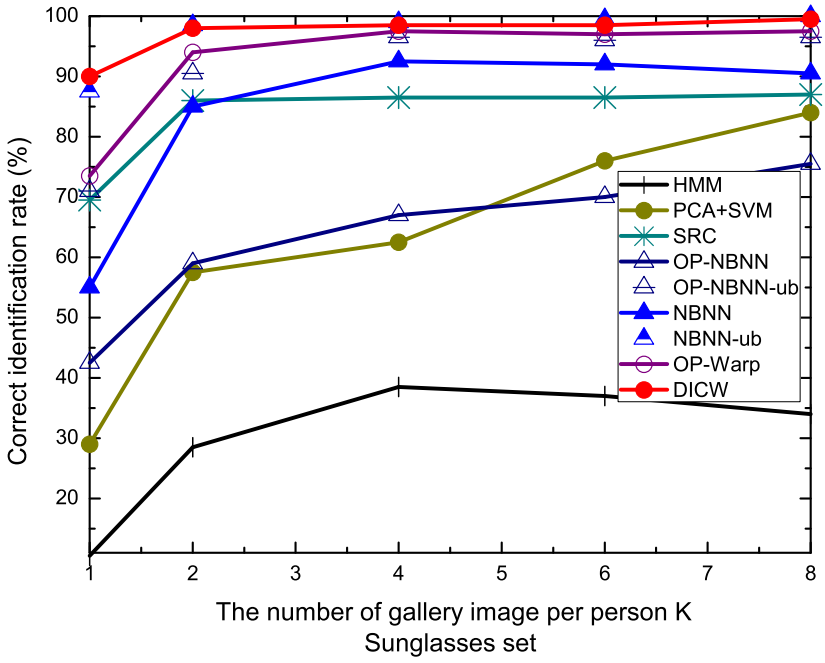
Uvs.O: The unoccluded frontal view images with various expressions are used as the gallery images (eight images per subject). For each subject, we select $K = 1, 2, 4, 6$ and 8 images to form the gallery sets, respectively. Two separate image sets (200 images each) containing sunglasses (cover about 30% of the image) and scarves (cover about 50% of the image) are used as probe sets, respectively. Fig. 4.11 shows the recognition results. The correct identification rates increase when more gallery images are available. HMM and SVM are generic training based methods and are unalbe to deal with *unseen* occlusions in

the probe images. In the scarf testing set, the performance of SRC deteriorates significantly compared with that on the sunglasses set due to the occluded area is much larger (i.e., not *sparse*). Local matching based NBNN and DICW perform better than others on the whole. With the empirically best values of location weights, NBNN-ub achieves very comparable performance to DICW. But DICW is slightly superior. Even at $K = 1$, DICW still achieves 90% and 83% on the sunglasses set and scarf set, respectively.

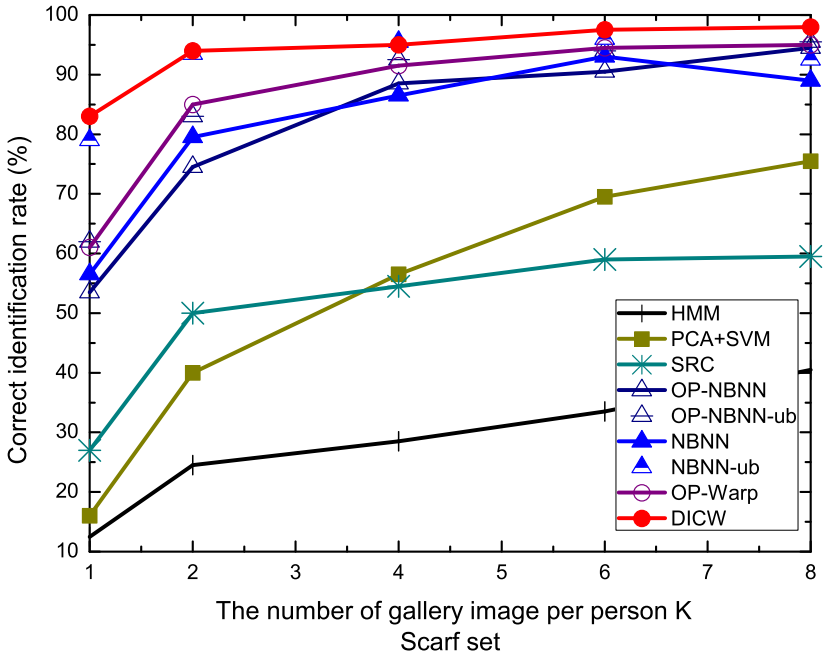
With the same experimental setting, we also compare DICW with the state-of-the-art algorithms (using eight gallery images per subject, $K = 8$). The results are shown in Table 4.2. Only the pixel intensity is used except the MLERPM method. MLERPM, which is also a local matching based method as ours, uses the SIFT [130] and SURF [164] features to handle the misalignment of images. Note that compared with other methods which are reconstruction based, our method does not require any data-dependent training. SRC-partition and CRC-RLS-partition indicate the strategy of partitioning an image into 4×2 local patches for performance improvement for the original method SRC [25] and CRC-RLS [119], respectively. Patch based methods deal with the occlusions locally and are more robust. DICW achieves comparable or better recognition rates among these methods and with a relatively low computational cost (see Section 4.5.6). In the scarf set, albeit the fact that nearly half of the face is occluded, only 2% images are misclassified by DICW. To the best of our knowledge, this is the best result achieved on the scarf set under the same experimental setting.

Table 4.2: **Uvs.O**: comparison of DICW and the-state-of-the-art methods (K=8)

Method	Sunglasses	Scarf	Average	Feature
SRC-partition [25]	97.5	93.5	95.5	Intensity
LRC [116]	96.0	26.0	61.0	
CRC-RLS-partition [119]	91.5	95.0	93.3	
SEC-MRF [114]	99.0∼**100**	95.0∼ 97.5	97.0∼**98.8**	
l_{struct} [151]	99.5	87.5	93.5	
OP-Warp	97.5	95.0	96.3	
Proposed DICW	99.5	**98.0**	**98.8**	
MLERPM [124]	98.0	97.0	97.5	SIFT [130] & SURF [164]



(a)



(b)

Figure 4.11: **Uvs.O**: identification results on the AR database with a) sunglasses occlusion and b) scarf occlusion.

Ovs.U and **Ovs.O**: For the **Ovs.U** scenario, we select four images with sunglasses and scarves (Figure 4.10b) to form the gallery set and eight unoccluded images as the probe set (Figure 4.10a). For the **Ovs.O** scenario, we conduct two experiments: 1) two images with scarves as the gallery set and two images with sunglasses as the probe set; 2) vice versa. Note that with this setting, in each test the occlusion type in the gallery set is *different* from that in the probe set, which is very challenging for recognition.

The results are shown in Fig. 4.12. On the gallery set which contains occluded faces, the results of HMM and SVM are much worse than others as expected. In the **Ovs.O** testing, there are only two gallery images per subject. It is very difficult for SRC to reconstruct an unoccluded probe image with such limited number of gallery images. Local matching based NBNN and our DICW perform better. Comparing OP-NBNN with OP-NBNN-ub, and NBNN with NBNN-ub, it can be found that the performance of NBNN is highly dependent on the empirically best values of location weights. Overall, DICW consistently outperforms the best of NBNN (i.e., NBNN-ub) by about 4% on average.

4.4.3 Face identification with general occlusions in realistic environments

In this Section, we test our method on the TFWM [135] database. In our experiments, we use images of 100 subjects (ten images per subject) containing various types of occlusions (Figure 2.11). For each subject, we choose $K = 1, 3, 5$ and 8 unoccluded images as gallery sets, respectively, and the remaining two images as the probe set. Occlusions occur at random in the gallery or probe set or in both. This includes all the three occlusion scenarios in Chapter 2. The face area of each image is cropped from the background and re-sized to 80×60 pixels and the patch size is 5×5 pixels. Only the pixel intensity is used in all methods.

The recognition results are shown in Fig. 4.13. Note that the images used in the experiments are *not well aligned* due to the variations. Some occlusions (e.g., hand) have very similar texture as the face, which are difficult to be detected by skin colour based models [33]. NBNN, which only relies on the texture similarity without considering the

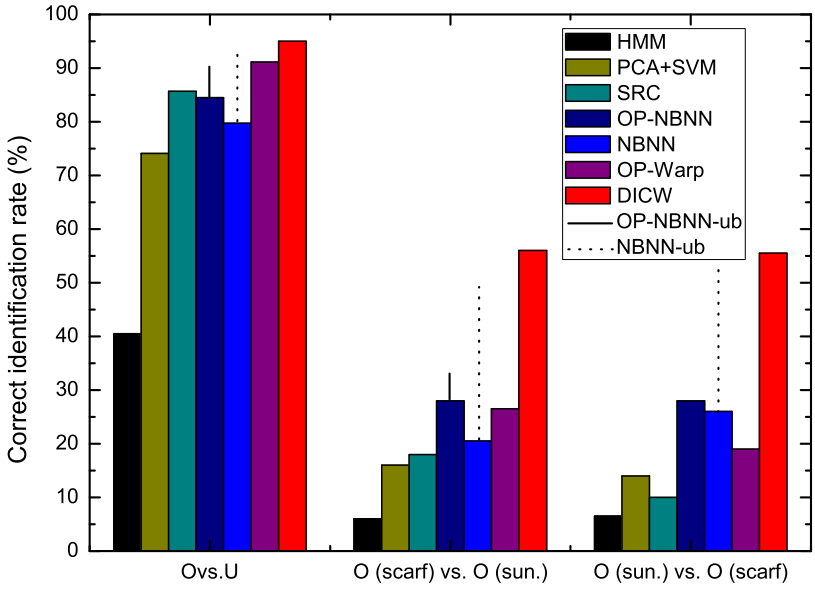


Figure 4.12: **Ovs.U** and **Ovs.O**: identification results on the AR database with occlusions in gallery or/and probe sets.

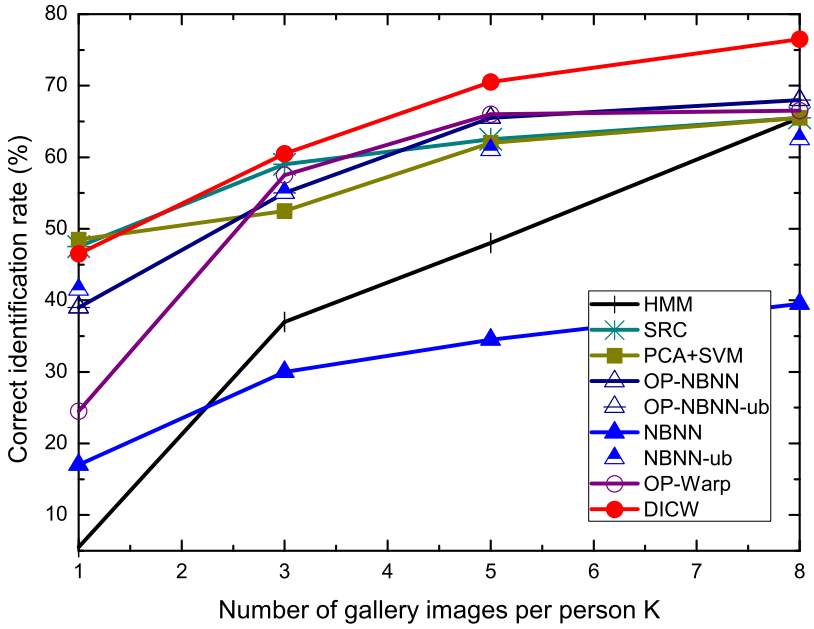


Figure 4.13: Identification results on the TFWM database.

structural constraint of a face, does not achieve comparable performance as ours. As more gallery images are available, the accuracies of all methods increase. When $K = 8$, most methods reach a *bottleneck* with the rate around 65%. DICW outperforms these methods by a notable margin.

4.5 Discussion

4.5.1 The effect of patch size

The size of the image patch can have significant impact on the performance. To investigate this factor, we use 400 unoccluded images (size of 80×65 pixels) of 100 subjects from the FRGC database as the gallery set and 400 images in each of six probe sets, which contain randomly located occlusions (as mentioned in Section 4.4.1) from 0% to 50% level, respectively. We test our DICW with the patch sizes from 3×3 pixels to 10×10 pixels. The

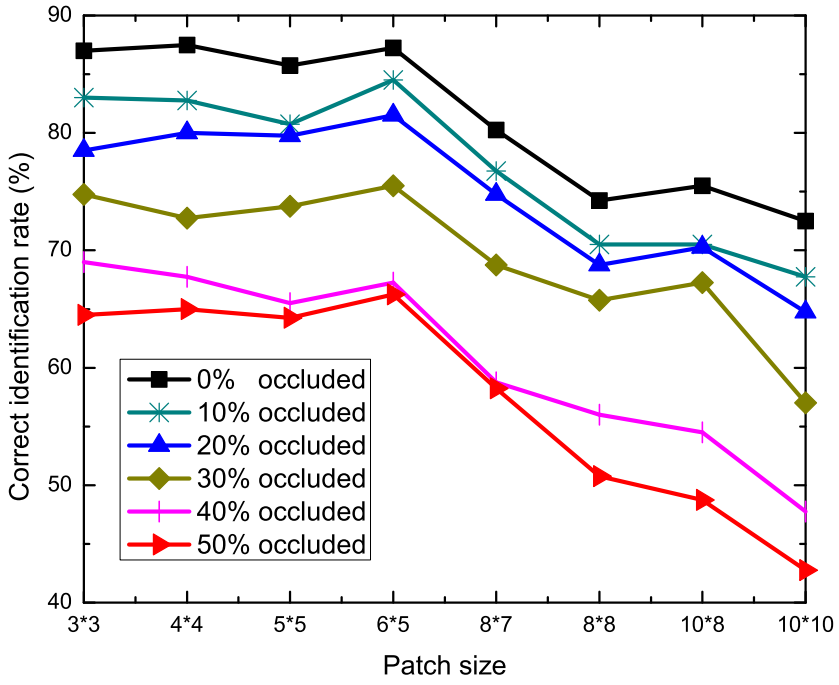


Figure 4.14: Correct identification rates with respect to the patch size.

correct identification rates with respect to the patch size are shown as Figure 4.14. There is no sharp fluctuation in each of the rate curve when the patch size is less than or equal to 6×5 pixels. Our method is robust to different patch sizes in an appropriate range despite the ratio of occlusions. The relatively smaller patches lead to better recognition rate since they provide more flexibility to use spatial information than the larger ones. Based on the experimental results, sizes smaller than 6×5 pixels are recommended.

When high resolution face images are available (i.e., larger size of the whole image), the number of patches will increase with the recommended patch sizes. In the experiments, we follow the most used image size (e.g., 80×65 pixels) in current works. More investigation should be done in the future to determine the best ratio of the patch size to the image size.

4.5.2 The effect of patch overlap

In the previous experiments we used the difference patch to enhance the textured features in patches. It is interesting to see if the overlapping patch has this similar effect. We conducted experiments on the AR database to investigate this since it contains real occlusions with different textures. We selected four unoccluded images from Session 1 for each subject as the gallery set and two images with sunglasses and scarves from Session 2 as the probe set so the testing dataset contains variations of occlusion and illumination changes. We tested the use of different patch sizes (4×4 to 16×15 pixels) with different overlap ratios (0%, 25%, 50%, 75%) and compared their results with that of using the difference patch. 25% ratio means the adjacent patches have a 25% horizontal overlap. So the larger the ratio is, the larger the number of patches will be in each image sequence. Note that 0% overlap ratio means using the original patches. The intensity values of each patch are used.

Fig. 4.15 shows the recognition results. Large overlap ratio leads to better accuracy. Note that higher overlap ratio also increases the number of patches in each image sequence, which leads to a higher computational cost. For small patch sizes (i.e., $4 \times 4, 5 \times 5$ and 8×6 pixels), using the difference patch yields significantly better results than using the overlapping patch. This is compatible with our analysis in Section 4.1. A difference patch is the approximation of the first-order derivative of adjacent *small* patches. The first-order derivative operator is sensitive to edges, which is able to enhance the textured regions. When the patch becomes large (i.e., $10 \times 10, 10 \times 15$ and 16×15 pixels), the advantage of using the difference patch is not obvious. This is reasonable since the texture in a large patch is less uniform. Note that the overall performance of using the small patch is better than that of using the large patch. Our DICW is compatible with the small patch as analysed in Section 4.5.1 so in the experiments we used the best one, the difference patch instead of the overlapping patch.

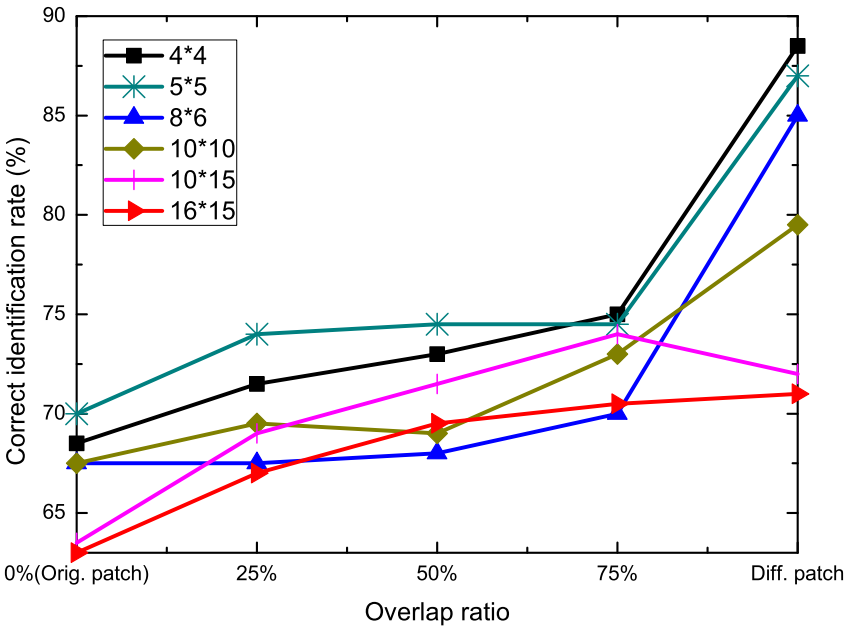


Figure 4.15: Identification rates (%) with respect to the overlap ratio comparing with using the difference patches.

4.5.3 The effect of image descriptor

In Section 4.5.2, our experiments indicate that the difference patch leads to better accuracy since it is able to enhance the textured regions in a face image. In this section we will carry out experiments to compare the discriminative power of the proposed difference patches and other local image descriptors such as LBP [50] and dense SIFT [165]. We use the same dataset in Section 4.5.2 and test both small patch size (i.e., 5×5 pixels) and large patch size (16×15 pixels).

Fig. 4.16 shows the recognition results. For large patch size, as we analysed before, the difference patch does not perform very well. For small patches, the performance of difference patch is comparable with that of SIFT and LBP. Note that the computation of difference patch is much simpler than other images descriptors. From Fig. 4.16 we can see, the local image descriptor is able to strengthen DICW when the image contains variations such as illumination changes and occlusions. When dealing with the uncontrollable data, applying these local features can further improve the performance of DICW. In the following experiments, LBP is adopted rather than the dense SIFT since LBP is computationally efficient

4.5.4 Robustness to misalignment

The face registration error can largely degrade the recognition performance [3] as we mentioned in Chapter 2. To evaluate the robustness of DICW to the misalignment of face images, similar to the work in [166], we use a subset of the AR database with 110 subjects (referred to AR-VJ). The faces in AR-VJ are automatically detected by the Viola & Jones detector [55] and cropped directly from the images without any alignment. Different from the images in the original AR database which are well cropped (Figure 4.9), these images contain large cropping and alignment errors as shown in Fig. 4.17. Some gallery images are even not correctly cropped (Fig. 4.17c), which can be seen as outliters in the gallery set.

Following the same experimental setting in [166], seven images of each subject from the first session are used as the gallery set and the other seven images from the second

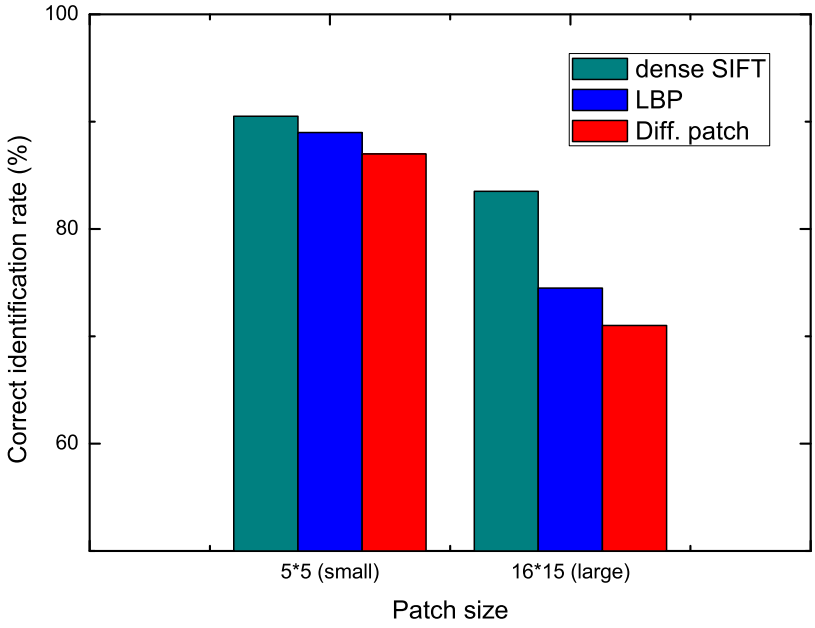


Figure 4.16: Identification rates (%) of using different image descriptors and the difference patches.

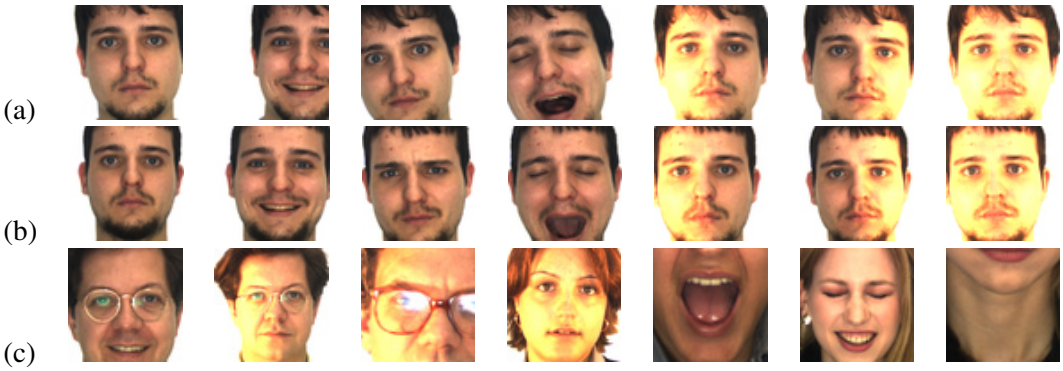


Figure 4.17: Sample images from the AR-VJ dataset without alignment. a) Gallery and b) probe images from the same subject. c) Gallery images with large cropping errors.

session as the probe set. All images are re-sized to 65×65 pixels and the patch size is 5×5 pixels. We use the $\text{LBP}^{u2}_{8,2}$ descriptor [50] for feature extraction to handle the illumination variations.

The recognition results are shown in Table 4.3. Our method outperforms other methods and achieves very close result with P2DW-FOSE [166], which is also a training-free method like ours. But different from our method, which performs warping on the *patch level*, P2DW-FOSE is a pseudo 2D warping method on the *pixel level* and its time complexity is quadratic in the number of pixels [166].

Table 4.3: Identification rates (%) on the AR-VJ dataset

Method	Correct identification rate (%)
Av-SpPCA[1][167]	93.6
DCT[1][3]	95.3
SURF-Face [168]	95.9
P2DW-FOSE [166]	**98.2**
Proposed DICW	97.3

[1] Using manually aligned images

4.5.5 The extension to face verification in the wild

In this Section, we extend DICW for face verification tasks using the LFW database [134], which is the most active benchmark for face recognition. The task of face verification under the LFW database's protocol is to determine if a pair of face images belongs to the same subject or not. Note that in the verification of each pair, it is a *Image-to-Image* comparison. So the experiments on the LFW database can be considered as an evaluation for the effectiveness of DICW when only *time warping* is used (no *within-class warping*).

Following the testing protocol of *View 2*, we use the most difficult experimental setting: *restricted unsupervised setting* where no class label information is available. In *View 2*, there are 3,000 *matched* (i.e., positive) and 3,000 *mismatched* (i.e., negative) image pairs. They are equally divided into ten randomly generated sets and the final verification performance is reported in terms of the mean classification accuracy over ten-fold cross-

validation. Here image pairs are classified into *the same subject* or *different subjects* by thresholding on their distance. Similar to the work in [161, 169–171], we use the LFW-a version provided at the LFW website[3] and $\text{LBP}^{u2}_{8,2}$ as feature descriptor. All images are cropped and re-sized to 150×80 pixels as suggested in [172] and the patch size is 3×3 pixels.

Chen *et al.*'s work [173] produces very competitive results on the LFW database by using the high-dimensional LBP feature. It is confirmed that features sampled at facial landmarks lead to better recognition performance than those sampled from regular grids. Motivated by this, we also select 25 landmarks [43] of the inner face and follow the similar process as in [173]: 1) normalise the unaligned images according to 2 facial landmarks (i.e., the tip of the nose and the centre of the mouth), and 2) extract image blocks (size of 30×30 pixels) centred around 25 facial landmarks from each image. Each block is partitioned into 3×3 pixels patches which are then concatenated to form a sequence. The original DICW algorithm is performed according to each block (i.e., sequence) and a corresponding distance is generated respectively. The sum of these distances is the final distance for each image pair. We refer our method with this strategy (i.e., sampling features around landmarks) as DICW-L and the original DICW (i.e., sampling features from regular grids) as DICW-G.

LFW is an extremely challenging database containing large variations, especially pose changes. As presented in [174], the first several principal components (PCs) usually capture these variations in the principal component analysis (PCA) subspace [38]. Therefore, we adopt the component analysis process in [174] to remove the first several PCs for performance according to:

$$\boldsymbol{F}' = \boldsymbol{F} - \boldsymbol{X}_i \boldsymbol{X}_i^T \boldsymbol{F} \tag{4.8}$$

where $\boldsymbol{F}$ is the original feature vector of an image by concatenating all the patch features of the image sequence (i.e., $\boldsymbol{P}$ or $\boldsymbol{G}_k$ in Section 4.2) and $\boldsymbol{X}_i$ is the first i components in the PCA subspace. We quantitatively test the value of i using the *View 1* dataset provided

[3] http://vis-www.cs.umass.edu/lfw/

by the LFW database and set the best value $i = 8$. $\boldsymbol{F}'$ is the improved feature vector used in the experiments for the LFW database. In this way, the large variations can be reduced to some extent. At the same time, different from the general dimension reduction operation (i.e., the original PCA), the topological structure of each image is still maintained so our patch based DICW can be performed directly on the improved features by this process.

We compare DICW with other methods under the same testing protocol *without outside training data*. In the experiments, only $\text{LBP}^{u2}_{8,2}$ descriptor [50] is used. We draw the the ROC (Receiver Operating Characteristic) curves of DICW and other state-of-the-art methods in Fig. 4.18. It shows the performance of DICW-G is better than other methods which use only single feature such as SD-MATCHES (SIFT [130]), H-XS-40 (LBP [50]), GJD-BC-100 (Gabor [131]), LARK (locally adaptive regression kernel descriptor [169]) and LHS (local higher-order statistics [170]). When extracting features around facial landmarks, the performance of DICW is further improved with a large margin. The area under the ROC curve (AUC) of DICW-L is 0.874 as shown in Table 4.4, which is the best among all methods. These experimental results confirm the effectiveness of DICW even when only *time warping* is performed.

Table 4.4: Area under the ROC curve (AUC) on the LFW database under unsupervised setting

Method	AUC	Feature extraction
SD-MATCHES [172]	0.5407	From grids
H-XS-40 [172]	0.7574	
GJD-BC-100 [172]	0.7392	
LARK [169]	0.7830	
LHS [170]	0.8107	
Proposed DICW-G	0.8286	
Proposed DICW-L	**0.8740**	From landmarks

4.5.6 The computational complexity and usability analysis

From Algorithm 2 in Section 4.3, we can see that the time complexity of DICW for computing the distance between a query image and an enrolled class is $\boldsymbol{O}(\max\{M, N\}lK)$,

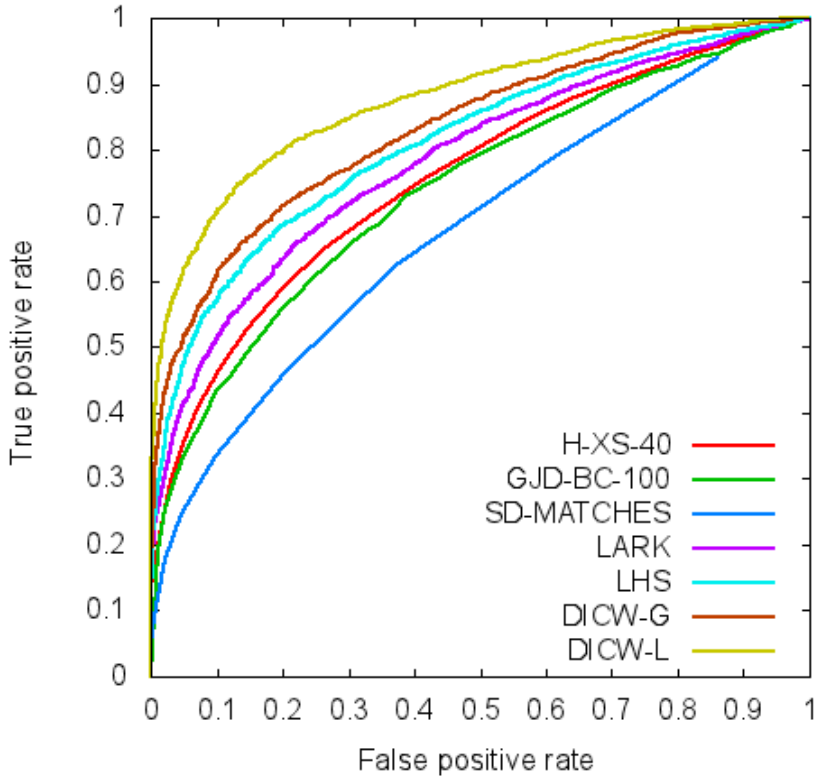


Figure 4.18: ROC curves of the-state-of-the-art methods and our DICW on the LFW database.

where M, N are the numbers of patches in each probe sequence and gallery sequence, respectively. l is the window width as mentioned in Section 4.2. For better readability, here we use M' to represent $\max\{M, N\}$. The number of gallery images per class K is very small compared with the number of patches M' in each sequence (i.e., $K \ll M'$). Thus the complexity is represented as $\boldsymbol{O}(M'l)$. Note that usually $l = 10\%M'$, so the warping distance can be obtained very efficiently. On the other hand, the computational cost of the reconstruction based method SRC is very high [25]. For an intuitive comparison, Table 4.5 shows the runtime of DICW and SRC [4] for classifying a query image under the same setting as the experiments of Table 4.2 (i.e., $M = N = 192, K = 8, u = 83 \times 60 = 4980$ and $v = 799$, 100 enrolled subjects in total[5]) using Matlab implementation (running on a platform with quad-core 3.10GHz CPUs and 8 GB memory). DICW is about 15 times faster than SRC [25] when classifying a query image.

Table 4.5: Comparison of average runtime (s)

	Per class	All class
SRC [25]	N/A	89
Proposed DICW	0.05	6

Compared with the reconstruction based approaches, which represent a query image using all enrolled images, DICW computes the distance between the probe image and each enrolled class independently. So in the real face recognition applications, the distance matrix can be generated in parallel and the enrolled database can be updated incrementally. This is very practical for real-world applications.

4.6 Further analysis and improvement

In the previous sections, we evaluated DICW through extensive experiments on face images with large variations. In this section we will further analysis why the DICW works compared with similar methods, and when and why it will fail.

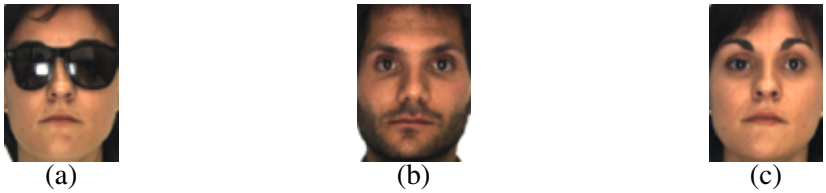
[4] We use the *l1_ls* package for implementation. `http://www.stanford.edu/~boyd/l1_ls/`

[5] A corrupted image w-027-14.bmp is excepted

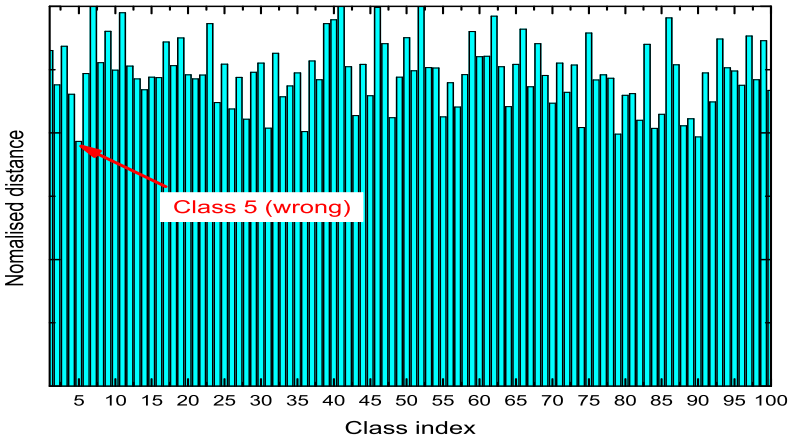
NBNN [163] presented in the previous sections is a similar method to ours. It also calculates the *Image-to-Class* distance between a probe patch set and a gallery patch set from a given class. The difference is that it does not consider the spatial relationship between patches like ours and each probe patch can be matched to any patches from any location in the gallery patch set. Fig. 4.19 is an illustration example. The occluded probe image is from Class 74 but is incorrectly classified to the Class 5 by NBNN. Actually the images from Class 74 and Class 5 are not alike. But the texture of sunglasses is very similar to that of beard in Class 5. Without the location constraint, the beard patches are wrongly matched to the sunglasses thus the distance is affected by this occlusion. On the other hand, DICW keeps the order information and matches patches within a proper range which leads to correct classification.

Fig. 4.20 shows the differences of NBNN and DICW when computing the *Image-to-Class* distance between the probe and the gallery. NBNN calculates the distance between two patch sets and the overall distance is the sum of patch-pair distances. On the other hand, in DICW, the probe set and gallery patch set are ordered. The spatial relationship between patches is encoded. When a probe patch is matched to a gallery patch, the following probe patches will only be matched to the gallery patches within a proper range. This is guaranteed by the four constraints mentioned in Section 4.2. DICW actually tries every possible combination of matching correspondence of patch pairs so the final matching is the global optimum for the probe patch set and the gallery patch set. Compared with NBNN, DICW considers both the texture similarity and the geometric similarity of patches. The work in [175] points out that the contextual information between facial features plays an important role in recognition. Our work confirms their observation. Although a location weight can be adopted in NBNN, the weight needs to be manually set for different testing dataset as analysed before, which is not suitable for practical applications. In DICW, the order constraint is naturally encoded during distance computation.

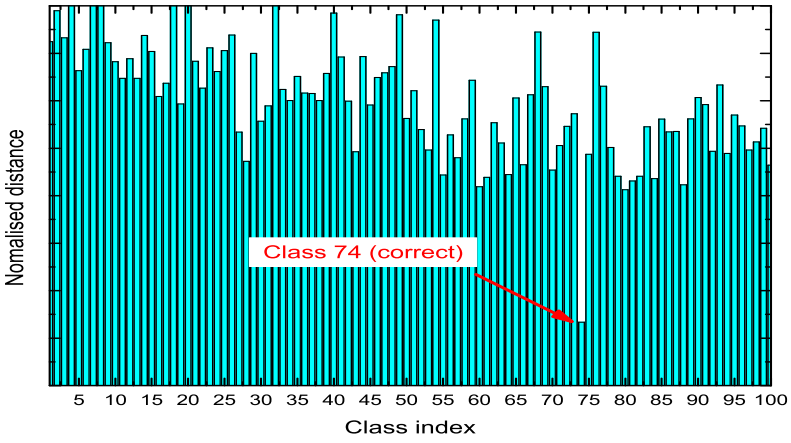
DICW represents a face image as a patch sequence which maintains the order of the facial features of the face. To some extent, the geometric information of a face



(a) (b) (c)



(d) By NBNN



(e) By DICW

Figure 4.19: Comparison of classification results by NBNN and DICW: a) The probe image from Class 74. b) Classification result (Class 5) by NBNN. c) Classification result (Class 74) by DICW. Distance to each class computed d) by NBNN and e) by DICW, respectively.

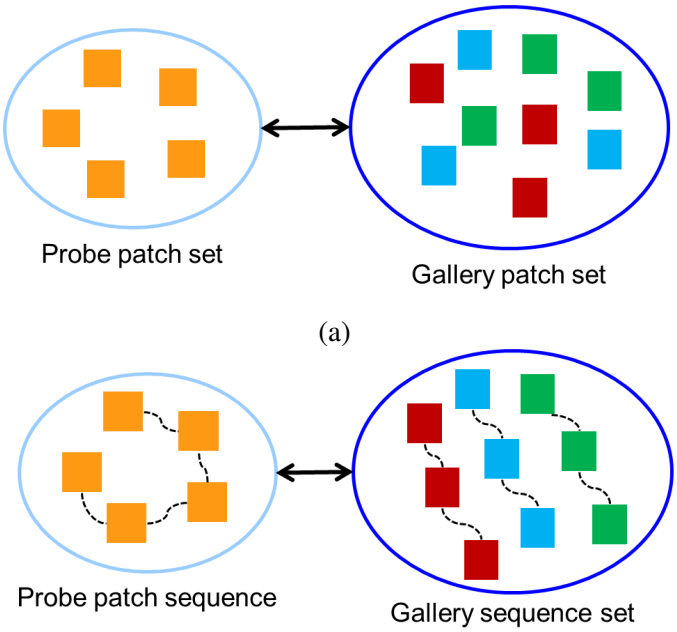


Figure 4.20: The difference of NBNN and DICW: a) NBNN: calculates the distance between two patch sets. b) DICW: calculates the distance between two ordered sequence patch set.

is reduced from 2D to 1D. However, the direct 2D image warping is an NP-complete problem [176]. P2DW-FOSE mentioned in Section 4.5.4 is a pseudo 2D warping method but with a remarkably large computational cost (i.e., a quadratic function of the number of pixels) [166]. DICW incurs a lower computational cost due to its patch sequence representation. In addition, each patch still contains the local 2D information which is helpful for classification.

Fig. 4.21 shows a failure example which can not be correctly classified by both DICW and NBNN. The discriminative eyes region is occluded by sunglasses, which makes recognition difficult. In addition, a probe face with sunglasses (Fig. 4.21a) is more similar to a gallery face with glasses (Fig. 4.21b) in the feature space, which leads to misclassification.

Looking back to the definition of DICW in Section 4.2, although *warping* is helpful for avoiding large distance error caused by occlusions, the occluded area is not directly removed during matching. Here we employ a simple but very effective scheme for

(a) (b) (c)

Figure 4.21: Failure example by DICW: a) a probe image from Class 51, b) the wrong class (Class 72) classified by DICW, and c) the gallery image from Class 51.

improving the performance of DICW. As shown in Fig. 4.22, we do not use all patches in a probe sequence for warping, instead, we randomly select a subset of patch set then compute the *Image-to-Class* distance based on this subset. We repeat this n times and generate a class label (the class with the shortest distance) each time according to the calculated distance. Finally, the final class label is decided by majority voting by n *experts*. With random selection, it is possible to skip the occluded patches. It is also possible that the occluded patches are chosen but this effect will be eliminated by the majority voting strategy since we assume that the occluded areas only take up a small part of a face. This assumption is reasonable since if most parts of a face are occluded, even a human being will feel difficult to recognise it. Different from the *occlusion detection* based methods which attempt to detect and remove occlusion area as we mentioned before, this simple strategy does not rely on any prior knowledge nor any data-dependent training.

Here we use the same setting to Section 4.5.2 (patch size: 5×5 pixels). We randomly select 15% patches in a sequence (30 patches from 192 patches) each time as an *expert* and select $n = 50$ *experts* in total. Since this scheme is based on random selection, we repeat the whole classification process ten times and calculate the average identification rate. The results are shown in Table 4.6. The performance of DICW is improved by 2% on average by using only 50 *experts* (Note that for each *expert*, the computation of DICW is much faster than before since the number of *subset* patches is much smaller than that of the whole sequence). Generally, involving more experts will lead to higher accuracy since this increases the diversity of decision *views*, which is more robust to different variations. But this will also raise the whole computational cost, which needs to be considered to keep

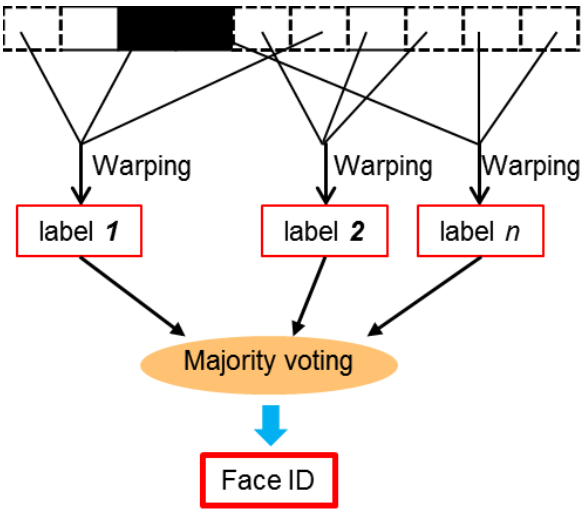


Figure 4.22: Random selection and majority voting scheme for improving the performance of DICW.

a balance between accuracy and computation. The improvement is more obvious when the number of image per class is limited. In next chapter, we will introduce the details of this scheme to improve the performance of DICW especially when $K = 1$.

Table 4.6: Identification rates (%) of DICW and the improvement scheme on the AR database

# Img./class (K)	1	2	3	4
DICW	81.0	83.5	86.0	87.0
Improvement scheme	84.5	85.2	86.5	89.0

4.7 Summary

Most of the existing occluded face recognition works that simply treat occluded face recognition as a recovery problem or just employ the framework for general object classification, neglect the inherent structure of the face. Wang *et al.* proposed a Markov Random Field (MRF) based method [175] for face recognition and confirmed that contextual information between facial features plays an important role in recognition. In this chapter, the proposed method DICW takes the *facial order*, which contains the geometry information of the face, into account when recognising partially occluded faces.

We first represent a face image as an ordered sequence, then treat the image matching problem as the process of finding optimal alignment between a probe sequence and a set of gallery sequences. Finally, we employ the dynamic programming technique to compute the *Image-to-Class* distance for classification. Extensive experiments on the FRGC, AR, TFWM and LFW face databases show that DICW achieves promising performance when handling various types of occlusions. In addition, in the most challenging cases where occlusions exist in both gallery and probe sets and only a limited number of gallery images are available for each subject, DICW still performs satisfactorily.

In uncontrollable environments with non-cooperative subjects, the occlusion pre-processing and the collection of sufficient and representative training samples are generally very difficult. Our DICW can be applied directly to face images without performing occlusion detection in advance and does not require a training process. All of these make our approach more applicable in real-world scenarios. Given its merits, DICW can be applied to deal with other problems caused by local deformations in face recognition (e.g., the facial expression problem), as well as other object recognition problems where the geometric relationship or contextual information of features should be considered.

Chapter 5

Extension of DICW: fixations and saccades based classification

As introduced in Chapter 4, DICW still works well with limited number of gallery images per subject compared with the reconstruction based methods presented in Chapter 3. In some real-world scenarios such as law enforcement, usually only one image is available for one subject. The performance of the traditional learning based methods [38, 39] will suffer because the training samples are limited. Many approaches are proposed to handle this *single sample per person* (SSPP) problem [18]. The work in [120] creates a suitable self-organising map (SOM) from images for representing each person. Partial distance (PD) [121] uses non-metric partial similarity measure for matching with a similarity threshold which is learned from the training set. Discriminative multi-manifold analysis (DMMA) [177] formulates the SSPP face recognition as a manifold-to-manifold matching problem by learning multiple feature spaces to maximise the manifold margins of different persons. These methods are more or less model-based. The thresholds or parameters in the model are trained on a representative data set.

Different from other biometric traits such as fingerprint and iris, the face is a natural trait for humans to recognise a person even met just once. Inspired by the observation of fixations and saccades in human visual perception, we propose a novel method - *fixations*

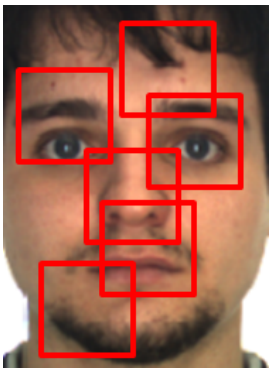


Figure 5.1: Illustration of fixations and saccades in human visual perception for a face. The red boxes indicate the fixations.

and saccades based classification (FSC), to solve the SSPP problem. FSC combines our previous DICW algorithm and the majority voting strategy. Besides occluded face images, in the experiments we also test FSC on images with large expression changes. Experimental results confirm that FSC is robust to local deformations of a face even only one gallery image is available for each person.

The rest of this chapter is organised as follows. Section 5.1 introduces the background of our method from the view of face recognition by humans. Section 5.2 explains our method in details. To evaluate the effectiveness of our method, extensive experiments are conducted on the FRGC and the AR databases and the results are reported in Section 5.3. Finally, Section 5.4 concludes the chapter. Note that symbols used are only valid within this chapter.

5.1 Background

The face is a natural trait for humans to recognise each other. Neuroscientists, psychologists and computer scientists all show interests in the research of face recognition. Although the mechanism of face recognition by humans is not fully understood, some observations on human visual perception can give us inspirations to design intelligent recognition algorithms.

When observing an object (e.g., face), humans do not look at it in fixed steadiness;

instead, the human eyes focus on small parts of the whole object and quickly switch between them [178]. The visual gaze on a single location is called *fixation* and a rapid movement of eyes is called a *saccade*. The term *fixation* also refers to a small part of focus rather than to the act of fixating. This is shown in Figure 5.1. One reason for the fixation and saccadic movement of eyes is: the central part of the retina which can provide the high resolution portion of vision is very small in humans, and small parts of an object can be scanned with greater resolution by moving eyes, which is helpful for recognition. Thus, when performing face recognition, humans quickly scan a set of *fixations* instead of the whole face [178]. For recognition algorithms, it is natural to simulate this process by sampling several small parts from a face image for feature extraction. In addition, considering the local deformations due to occlusions, the affected areas of a face are not helpful for recognition. Since the locations of deformations are unpredictable, random sampling is a good choice [179].

On the other hand, in human face recognition, the features are locally sampled by fixations but the whole facial structure is also considered [32]. One local fixation may not be sufficient for recognition, but a large field of random selections will be likely to produce a global, good output[1]. The information of a large number of fixations can be combined as a whole by a multiple classifier system. The combination can be implemented using a variety of strategies, among which majority voting is by far the simplest, and yet it has been found to be just as effective as more complicated schemes in improving the recognition results. In the next section we will introduce our method in details.

5.2 Fixations and saccades based classification

We introduced the DICW method in Chapter 4. In face recognition by humans, the structure of the face (i.e., the spatial relationship between facial features) is very important. Our DICW represents a face image as a patch sequence which conforms to the contextual order of the facial features. It employs both the local (i.e., patch-based features) and the global

[1]This is what is called the *law of large numbers* (LLN)

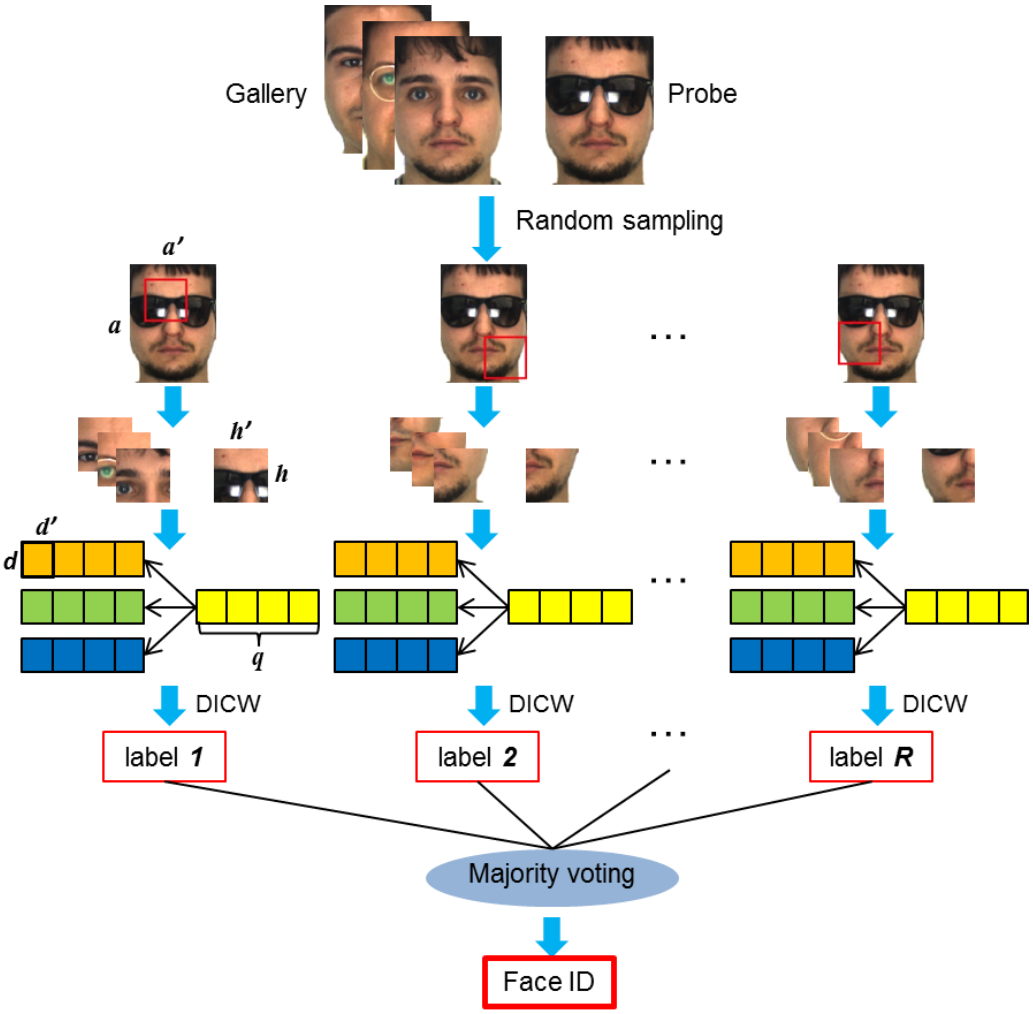


Figure 5.2: The framework of the FSC. $a \times a'$ is the size of an image. $h \times h'$ is the size of a fixation. $d \times d'$ is the size of a patch. q is the length of a patch sequence.

(i.e. facial order) information, which is compatible with the process of face recognition by humans.

Inspired by the aforementioned observations in human visual perception, we propose the FSC method which contains the following processes:

1. Random sampling of a large number of *fixations* from face images;

2. Performing DICW to calculate the distance between fixations and assigning a class label for each fixation;

3. Combining the classification decisions (i.e., labels) of all fixations using majority voting and making the final decision on the class of a face.

Each fixation can be seen as an *expert* for classifying a face. Looking back to the SSPP problem we mentioned at the beginning of this chapter, even only one single gallery image is available for each enrolled subject, a large number of *experts* can be helpful for classifying a face from their different *perspectives*.

On the other hand, let us consider the occlusions in face images. In the DICW presented in Chapter 4, all patches (occluded or non-occluded) are used during matching (i.e., *warping*). Occluded patches are not directly removed since the occluded locations are difficult to predict as we analysed in Chapter 2. With random sampling, it is possible to skip the occluded areas. It is also possible that the occluded areas are chosen but this effect will be eliminated by the majority voting strategy since we assume that the occluded areas only take up a small part of a face. This assumption is reasonable since if most parts of a face are occluded, even a human being will have difficulty recognising it. Note that compared with most of the works for occluded face recognition, our strategy does not rely on any prior knowledge. It is simple, yet effective.

A number of R fixations, $\{\boldsymbol{x}^1, ..., \boldsymbol{x}^r, ..., \boldsymbol{x}^R\}$, are randomly sampled from a face image. Each fixation (size of $e = h \times h'$ pixels) is partitioned into q (i.e., set $M = q$ in Section 4.2) patches (size of $s = d \times d'$ pixels) and then forms a sequence which maintains the contextual order. Then each fixation sequence is compared with the fixation sequences from the corresponding area of enrolled face images by DICW and a class label is generated

for each fixation. We define a binary function $f(r,l)$ for recording the voting result for each fixation as:

$$f(r,l) = \begin{cases} 1 & if \quad class(\boldsymbol{x}^r) = l \\ 0 & otherwise \end{cases} \tag{5.1}$$

where $l \in [1,2,...,L]$ and L is the number of classes. $class(\boldsymbol{x}^r)$ is the label of fixation $\boldsymbol{x}^r$ according to the DICW distance. Even just only one image is available per person, the final classification decision can be made by the majority voting of a large number of fixations:

$$assign \quad \boldsymbol{P} \rightarrow class\, l \quad if \quad \sum_{r=1}^{R} f(r,l) = \max_{i=1}^{L} \sum_{r=1}^{R} f(r,i) \tag{5.2}$$

where $\boldsymbol{P}$ is the probe image.

Majority voting does not assume prior knowledge of the behaviour of the individual classifier. Here each fixation has the possibility to classify the face correctly or wrongly. A final correct decision is made when the consensus is correct. The combined decision is wrong *only if* a majority of the fixations votes are wrong and they all make the *same* mis-classification. But this does not often happen due to the large number of different possible mis-classifications.

The framework of our FSC is shown in Figure 5.2 and the processes are summarised in Algorithm 4. As mentioned in Section 4.5.6, the time complexity of DICW is a quadratic function of the number of patches (i.e., q) in a sequence (here the complexity is $\boldsymbol{O}(q^2)$). So the time complexity of FSC is $\boldsymbol{O}(R(\frac{e}{s})^2)$ where $q = \frac{e}{s}$. e is the size of a fixation and s is the size of a patch in a fixation. R is the number of fixations extracted from one face image.

5.3 Experimental analysis

In this section, the FRGC database [52] and AR database [132] are used to evaluate the effectiveness of the proposed method, FSC. Experiments of face recognition with different occlusions (e.g., randomly located squares, sunglasses, scarves) are conducted on these

Algorithm 4 Fixations and saccades based classification

Input:

$\boldsymbol{G}$: the set of gallery images from L classes;
$\boldsymbol{P}$: a probe image;
R: the number of fixations sampled from an image;
q: the number of patches in a fixation sequence;

Output:

class: the class label of $\boldsymbol{P}$;

Set the record matrix $f[1:R, 1:L] = 0$;
Crop R fixations $\{\boldsymbol{x}^1, ..., \boldsymbol{x}^r, ..., \boldsymbol{x}^R\}$ at R random locations from the probe image $\boldsymbol{P}$ and partition each fixation into q patches to form a q-length sequence;
For each gallery image in $\boldsymbol{G}$, crop R fixations from the corresponding locations to $\boldsymbol{P}$ and partition each fixation into q patches to form a q-length sequence, respectively;
for $r = 1$ to R **do**
    Calculate the DICW distance between $\boldsymbol{x}^r$ from $\boldsymbol{P}$ and $\boldsymbol{x}^r$ from each image in $\boldsymbol{G}$ using Algorithm 2 in Chapter 4;
    Label $\boldsymbol{x}^r$ from $\boldsymbol{P}$ as the gallery class with the shortest DICW distance;
    if the label of $\boldsymbol{x}^r$ is l **then**
        $f[r, l] = 1$;
    end if
end for
if $\sum_{r=1}^{R} f(r, l) = \max_{i=1}^{L} \sum_{r=1}^{R} f(r, i)$ **then**
    $class(\boldsymbol{P}) = l$;
end if
return $class(\boldsymbol{P})$;

public databases. Moreover, we also test FSC on images with expressions (e.g., smile, anger, scream) since expression also cause deformations of a face. For feature extraction, we use LBP ($\text{LBP}^{u2}_{8,2}$) descriptor [50], which is insensitive to illumination changes and robust to small misalignment.

We quantitatively compare FSC with the methods mentioned at the beginning of this chapter as well as some methods based on similar ideas as ours. We set the fixation size e to 3% of an image, and use 300 fixations ($R = 300$) which will be about 1.5 minutes of viewing time for a face assuming three fixations per second [179]. The setting of parameters will be discussed in Section 5.3.4. We follow the settings in Section 4.4 and use patch size of $s = 6 \times 5$ pixels in the FRGC database and $s = 5 \times 5$ pixels in the AR database for each fixation. In all experiments, we run our method ten times and report the average identification rate.

5.3.1 Face identification with randomly located occlusions

We first evaluate FSC on the FRGC database with randomly located occlusions as introduced in Section 4.4.1. For each subject, we choose *one* image as the gallery set and four images as the probe set. As shown in Figure 4.6, in some cases most salient facial features are occluded, which is very challenging for recognition.

Table 5.1: Identification rates (%) of FSC on the FRGC database with single image per person

Occlusion	0%	10%	20%	30%	40%	50%
PCA+SVM [162]	69.5	65.8	57.3	36.8	36.8	22.3
SRC-partition [25]	65.8	55.8	47.8	39.8	32.5	22.8
DICW	79.3	77.8	77.3	72.8	70.8	64.5
Proposed FSC	**84.2**	**82.4**	**80.3**	**78.1**	**73.2**	**69.6**

There are six probe sets, each corresponding to a different level of occlusion (i.e., 0%, 10%, 20%, 30%, 40% 50%). The recognition results are shown in Table 5.1. As expected, the identification rates decrease when the level of occlusion increases. Our FSC outperforms DICW by about 5% and other methods such as the reconstruction based SRC

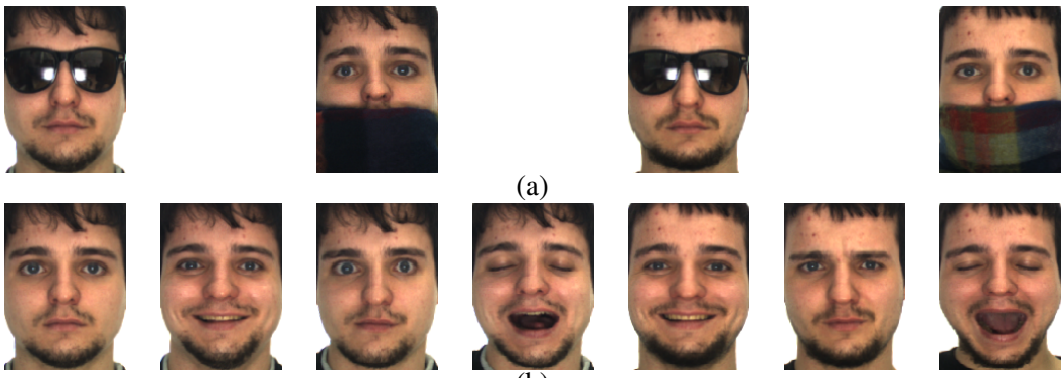


Figure 5.3: a) Cropped images from the AR database with occlusions. b) Cropped images from the AR database with different expressions.

[25] (using 4×2 block partitioning for performance improvement) and the supervised linear SVM [162] using PCA [38] for feature extraction (PCA+SVM). Even when half of the face is occluded, FSC still archives nearly 70% accuracy, which is much better than other methods.

5.3.2 Face identification with facial disguises

Next we investigate the robustness of FSC using partially occluded faces by real disguises. Similar to the works in [16, 17, 25, 120–122], in the experiments we choose a subset (50 male and 50 female subjects) of the AR database. For each subject, the neutral expression face from Session 1 is selected as the gallery set. The faces with sunglasses and scarf from both sessions are used as the probe sets (Figure 5.3a).

A comparison of the recognition results between FSC and other state-of-the-art methods is provided in Table 5.2. The performance of DICW is improved by the majority voting scheme, especially for the scarf set of session 2 (the most difficult set for other approaches), from 81.0% to 94.9%. Stringface [122] represents a face as a string of line segments, which also maintains the structural information of a face as DICW and the proposed method. FARO [180] is also a patch-based method as ours but is based on the partitioned iterated function system (PIFSs). SRC-partition [25], PD [121] and SOM [120] were introduced earlier. In $\text{PWCM}_{0.5}$ [16], an occlusion mask is trained through the use

of the skin colour. FSC clearly outperforms these approaches without any data-dependent training. CTSDP [181] is a 2D warping method which is also model-free, like ours. Its performance is improved by learning a suitable occlusion handling threshold on occluded images. The overall identification rate of FSC (96.6%) without occlusion pre-processing is very close to that of CTSDP (98.5%) with the occlusion threshold.

Table 5.2: Identification rates (%) on the AR database (occlusion) with single image per person

Method	Session 1		Session 2		Average	M/H[1]
	Sunglasses	Scarf	Sunglasses	Scarf		
Stringface [122]	88.0	96.0	76.0	88.0	87.0	No
FARO [180]	90.0	85.0	-	-	87.5	No
SRC-partition [25]	86.0	87.0	49.0	70.0	73.0	No
PD [121]	98.0	90.0	-	-	94.0	No
SOM [120]	97.0	95.0	60.0	52.0	76.0	No
CTSDP [181]	-	-	-	-	90.6	No
DICW	**99.0**	97.0	93.0	81.0	92.5	No
$PWCM_{0.5}$ [16]	97.0	94.0	72.0	71.0	83.5	Yes
CTSDP [181]	-	-	-	-	**98.5**	Yes
Proposed FSC	**99.0**	**98.7**	**93.7**	**94.9**	96.6	No

[1] Occlusion mask/threshold training required

5.3.3 Face identification with various expressions

We have evaluated the effectiveness of FSC using images with expression changes in the AR database. We use the same gallery set as in Section 5.3.2 and the images from both sessions with smile, anger and scream expressions as the probe sets (Figure 5.3b). The recognition results are shown in Table 5.3.

Most of the methods achieve satisfactory results on the testing images with smile and anger expressions from session 1 since these two expressions do not distort the face largely. Images from session 2 are more challenging because beside expressions, they also contain other variations such as different hair styles and illuminations. FSC outperforms the seven listed approaches in most cases. The scream expression causes large deformations of the face. The overall performance of all approaches on the scream sets (especially from

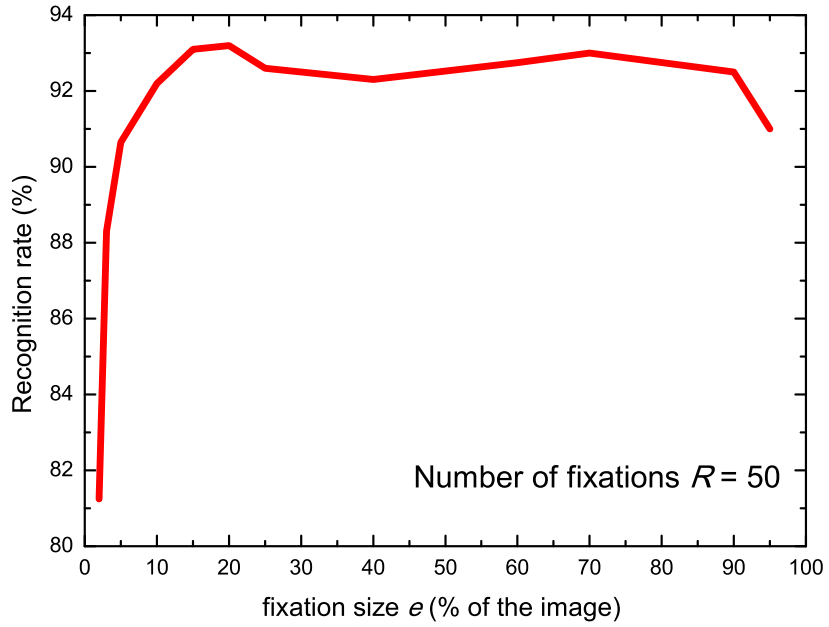
session 2) is relatively low due to its challenging nature. On the other hand, FSC achieves comparable rates with the 2D warping based method CTSDP. Note that the time complexity of CTSDP is $\boldsymbol{O}(i^2)$ [166] where $i = a \times a'$ (pixels) is the size of the image, compared with ours is just $\boldsymbol{O}(R(\frac{e}{s})^2)$. Here R, the number of fixations, can be seen as a constant. e is the fixation size and s is the size of the patch in a fixation sequence. Generally $e \ll i$ and $s > 1$. So FSC achieves comparable performance in most cases but is more efficient than CTSDP.

Table 5.3: Identification rate (%) on the AR database (expression) with single image per person

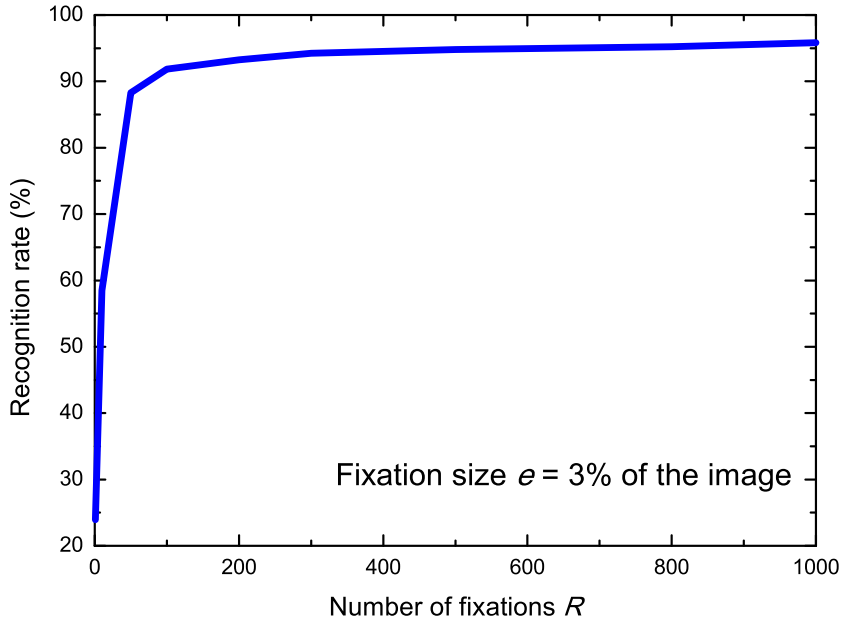
Method	Session 1			Session 2			Average
	Smile	Anger	Scream	Smile	Anger	Scream	
Stringface [122]	87.5	87.5	25.9	-	-	-	67.0
FARO [180]	96.0	-	60.0	-	-	-	78.0
SRC [25]	98.0	89.0	55.0	79.0	78.0	31.0	71.7
PD [121]	**100.0**	97.0	93.0	88.0	86.0	63.0	87.8
SOM [120]	**100.0**	98.0	88.0	88.0	90.0	64.0	88.0
DMMA [177]	99.0	93.0	69.0	85.0	79.0	45.0	78.3
DICW	**100.0**	99.0	84.0	91.0	92.0	44.0	85.0
CTSDP [181]	**100.0**	**100.0**	**95.5**	**98.2**	**99.1**	**86.4**	**96.5**
Proposed FSC	**100.0**	**100.0**	91.4	94.5	98.0	58.6	90.4

5.3.4 Discussion

The effect of patch size s on the recognition performance is discussed in Chapter 4. Here we fix s according to the settings in Section 4.5.1 and study the influence of the fixation size e and the number of fixations R. We conduct experiments on the AR database using images with sunglasses and scarves. The identification rates as a function of e and R when one parameter is fixed are shown in Figure 5.4. Intuitively, if e is too small ($< 3\%$ of the image), the order information contained in the fixation sequence will be very limited and not suitable for recognition. On the other hand, as mentioned in Section 5.2, since the time complexity of our method is $\boldsymbol{O}(R(\frac{e}{s})^2)$, the increase of e leads to higher computational cost (s is fixed). It can be seen from Figure 5.4b, the recognition rate is monotonically



(a)



(b)

Figure 5.4: a) Identification rate (%) as a function of the fixation size e and b) identification rate (%) as a function of the number of fixations R.

increasing with respect to the increasing R. Considering the computational efficiency, in our experiments, we set $e = 3\%$ of the image and empirically increase the number of fixations to $R = 300$ in order to gain higher recognition accuracy.

5.4 Summary

Inspired by the observation on fixations and saccades in human visual perception, we proposed a novel method FSC for the single image per person problem by combining our previous algorithm DICW with the majority voting strategy. On the two well-known face databases (FRGC and AR), FSC clearly outperforms the current approaches when dealing with occlusions and expression changes. In some extreme cases where the variations cause large deformations of the face, our method achieves comparable performance with the 2D warping based method at a much lower computational cost.

Chapter 6

Conclusions

The face is one of the most popular biometric traits used in the daily life for human identification. The past decades have seen significant progress in automatic face recognition (AFR). But the performance of the AFR techniques on images collected in unconstrained environments is still unsatisfactory. Uncontrollable illumination, pose, expression changes and occlusions pose acute challenges to face recognition techniques. Especially, the occlusion problem is relatively less studied compared with the problems of other uncontrollable factors.

In this thesis we focus on the recognition of faces with occlusions in unconstrained environments. We carefully analysed the occlusion problem and summarised the challenges to this problem. We dealt with the occlusion problem in two directions and proposed three novel algorithms to handle the occlusions in face images while also considering other factors. Our proposed algorithms were evaluated on standard face databases with various types of occlusions (e.g., randomly located occlusions, shadows, facial accessories, hands, hair) and experimental results confirmed their effectiveness. We also discussed several important and practical problems in face recognition (e.g., coupled variations, occlusions in gallery or/and probe sets and the SSPP problem [18]) and provide solutions to them.

6.1 Contributions and conclusions

We summarise our main contributions and the important research results as follows, which is helpful for the future research in face recognition and related areas.

1. In Chapter 2, we have provided a detailed literature review on the state-of-the-art face recognition methods for the occlusion problem. In real-world scenarios, the presence of occlusions is unpredictable. The three classical occlusion cases - **Uvs.O**, **Ovs.U** and **Ovs.O** should be considered when designing a real-world face recognition system. In addition, when conducting experiments to evaluate the effectiveness of an occluded face recognition algorithm, at least three kinds of occluded images should be considered: 1) images containing real occlusions such as disguises with various textures and shapes, 2) images containing randomly located occlusions without any prior knowledge of the location, and 3) images taken in natural conditions (e.g., outdoor environments) in which occlusions are coupled with other distorting factors.

2. In Chapter 3, we proposed a reconstruction based method *structured sparse representation (SSR) based face recognition* to simultaneously deal with the coupled condition of large illumination changes and occlusions. We proposed a structured occlusion dictionary for better modelling contiguous occlusions and employed an illumination insensitive WLD feature for handling severe illumination variations. Experimental results showed that the models considering *structured sparsity* can represent face images with occlusions better than the models considering *flat sparsity* [25]. The work in [25] showed that within the same SRC model, the classification accuracies of using different *holistic* features (i.e., Eigenface [38], Fisherface [39], Laplacianface [158], random projection and downsampled images) are very close. Our results showed the performance of reconstruction based methods such as SRC and SSR can be significantly improved by employing *local* features like WLD.

3. In Chapter 4, we proposed a local matching based method, *Dynamic Image-to-Class Warping* (DICW), as well as a face representation method: difference patch

sequence. DICW outperforms the-state-of-the-art algorithms when occlusions exist in gallery or/and probe images, especially when the number of gallery images for each subject is limited. To the best of our knowledge, DICW achieves the best identification rate in the current literature on the scarf set in the AR database with the same experimental setting. Moreover, DICW is robust to misalignment and incurs a lower computational cost compared with the reconstruction based methods. On the whole, local matching based methods performs better than reconstruction based methods when the gallery set is contaminated with occlusions. Our experimental results also suggested that the order of the facial features (i.e., the inherent structure of the face) is critical for recognition. On the other hand, using the difference patches for face representation archives much better results than using the original patches. Even with simper computation, the difference patches achieves comparable or better performance than other image features such as LBP and SIFT.

4. In Chapter 5, we proposed a novel method *fixations and saccades based classification* (FSC) which is an extension of DICW. We analysed some observations in human visual perception and simulated the process in these observations to improve the performance of DICW, especially for the SSPP problem. Even without occlusion mask/threshold training, the proposed FSC significantly outperforms the state-of-the-art algorithms which are able to handle the SSPP problem. It improves the accuracy of DICW by about 5% when tested on face images with randomly located occlusions and real disguises, respectively. Especially for the scarf set, which is the most difficult occlusion testing set, it improves DICW by nearly 14%. In addition, FSC outperforms other similar methods on testing sets with expression changes (i.e., smile, anger and scream) when only one neutral expression image is available in the gallery.

6.2 Future research directions

This thesis is just a brick we have contributed to the scientific community. A multitude of new research initiatives in the related areas are to be taken. Some possible lines of

investigations are as follows.

6.2.1 Investigations in the short-term

1. **Exploring the use of DICW and FSC algorithms for handling other variations which also cause local distortion of the face:** The (local) make-up (e.g., mascara, lipstick, rouge) [6, 7] and local plastic surgery (e.g., brow lift, blepharoplasty, rhinoplasty) [8–10] also affect the face appearance locally and are difficult to predict, like the occlusions. Collecting sufficient training samples is also not easy since these factors in testing images may be largely different from those in the training data. The local matching based DICW and FSC have the potential to handle these variations since the order of the facial features is maintained. More investigations of the changes on the skin texture and face components should be performed in the future.

2. **Investigating the features which are less sensitive or even invariant to the occlusion effect:** The work of this thesis focus on improving the robustness of face recognition algorithm to occlusions during the matching stage. As mentioned in Chapter 2, some recent methods [128, 129] also attempt to extract stable and occlusion-insensitive features from face images to handle the occlusion problem. We have proposed the simple but effective *difference patch* in Chapter 4 to represent occluded face images. The experimental results confirmed that its performance is comparable to or even better than some traditional image features. The difference patch is a preliminary study for investigating more sophisticated features to contend with occlusions. Using occlusion-insensitive features is able to further improve the performance of our proposed DICW and FSC as well as other matching algorithms for handling occlusions. This will be another future direction for our work.

6.2.2 Investigations in the long-term

1. **Combining face with other biometric traits for improving human identification accuracy:** In unconstrained environments, the face is not the only trait used by humans

to recognise each other. It is natural to combine face and other biometric traits such as gait, hair, head, etc. to improve the recognition performance [182]. How to represent the features from different modalities and how to fuse these features (e.g., feature level, score level, or decision level fusion) will be the important issues for future investigations.

2. **Combining face and other biometric traits for anti-spoofing:** Currently most works on multiple biometrics fusion focus on improving the identification accuracy of biometric systems. However, very few works consider the multiple biometrics fusion from the perspective of biometrics spoofing. Some researchers propose the idea of fusing face and voice for anti-spoofing [183]. Compared with that, fusing face and gait is more attractive and practical since it is more natural and these two modalities can be captured using the same type of sensor (e.g., camera, CCTV).

Besides the areas of pattern recognition, attacking face recognition algorithms using occlusions could be very interesting from the perspective of security. It is possible to adopt the knowledge and experience in dealing with the above problems in some novel and interesting ways to link the occlusion problem to the security related problems.

Bibliography

[1] P. Phillips, W. Scruggs, A. O'Toole, P. Flynn, K. Bowyer, C. Schott, and M. Sharpe, "FRVT 2006 and ICE 2006 large-scale experimental results," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 5, pp. 831 – 846, 2010.

[2] A. M. Martínez, "Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 6, pp. 748 – 763, Jun. 2002.

[3] H. K. Ekenel and R. Stiefelhagen, "Why is facial occlusion a challenging problem?" in *International Conference Biometrics (ICB)*, 2009, pp. 299 – 308.

[4] J. C. Klontz and A. K. Jain, "A case study on unconstrained facial recognition using the boston marathon bombings suspects," *Technical Report MSU-CSE-13-4*, 2013.

[5] S. Liao, A. K. Jain, and S. Z. Li, "Partial face recognition: Alignment-free approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 5, pp. 1193 – 1205, 2013.

[6] A. Dantcheva, C. Chen, and A. Ross, "Can facial cosmetics affect the matching accuracy of face recognition systems?" in *IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS)*, Sep. 2012, pp. 391 – 398.

[7] C. Chen, A. Dantcheva, and A. Ross, "Automatic facial makeup detection with application in face recognition," in *International Conference Biometrics (ICB)*, Jun. 2013, pp. 1 – 8.

[8] R. Singh, M. Vatsa, H. Bhatt, S. Bharadwaj, A. Noore, and S. Nooreyezdan, "Plastic surgery: A new dimension to face recognition," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 3, pp. 441 – 448, Sept 2010.

[9] G. Aggarwal, S. Biswas, P. J. Flynn, and K. W. Bowyer, "A sparse representation approach to face matching across plastic surgery," in *IEEE Workshop on the Applications of Computer Vision (WACV)*, 2012, pp. 113 – 119.

[10] X. Liu, S. Shan, and X. Chen, "Face recognition after plastic surgery: A comprehensive study," in *Asian Conference on Computer Vision (ACCV)*, ser. Lecture Notes in Computer Science, K. Lee, Y. Matsushita, J. Rehg, and Z. Hu, Eds. Springer Berlin Heidelberg, 2013, vol. 7725, pp. 565 – 576.

[11] M. Storer, M. Urschler, and H. Bischof, "Occlusion detection for ICAO compliant facial photographs," in *IEEE Conference Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2010, pp. 122 – 129.

[12] D. Lin and X. Tang, "Quality-driven face occlusion detection and recovery," in *IEEE Conference Computer Vision and Pattern Recognition (CVPR)*, 2007, pp. 1 – 7.

[13] T. Hosoi, S. Nagashima, K. Kobayashi, K. Ito, and T. Aoki, "Restoring occluded regions using FW-PCA for face recognition," in *IEEE Conference Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2012, pp. 23 – 30.

[14] R. Min, A. Hadid, and J.-L. Dugelay, "Improving the recognition of faces occluded by facial accessories," in *IEEE International Conference Automatic Face and Gesture Recognition (FG)*, 2011, pp. 442 – 447.

[15] R. Min, "Face recognition robust to occlusions," PhD Thesis, T*él*ecom Paris Tech, France, Apr. 2013.

[16] H. Jia and A. M. Martínez, "Face recognition with occlusions in the training

and testing sets," in *IEEE International Conference Automatic Face and Gesture Recognition (FG)*, Sep. 2008, pp. 1 – 6.

[17] H. Jia and A. M. Martínez, "Support vector machines in face recognition with occlusions," in *IEEE Conference Computer Vision and Pattern Recognition (CVPR)*, Jun. 2009, pp. 136 – 141.

[18] X. Tan, S. Chen, Z.-H. Zhou, and F. Zhang, "Face recognition from a single image per person: A survey," *Pattern Recognition*, vol. 39, no. 9, pp. 1725 – 1745, 2006.

[19] R. Reiter, "Readings in nonmonotonic reasoning," M. L. Ginsberg, Ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1987, ch. On Closed World Data Bases, pp. 300 – 310.

[20] X. Wei, C.-T. Li, and Y. Hu, "Robust face recognition under varying illumination and occlusion considering structured sparsity," in *International Conference Digital Image Computing Techniques and Applications (DICTA)*, 2012, pp. 1 – 7.

[21] X. Wei, C.-T. Li, and Y. Hu, "Face recognition with occlusion using dynamic image-to-class warping (DICW)," in *IEEE International Conference Automatic Face and Gesture Recognition (FG)*, 2013, pp. 1 – 6.

[22] X. Wei, C.-T. Li, and Y. Hu, "Robust face recognition with occlusions in both reference and query images," in *International Workshop Biometrics and Forensics*, 2013, pp. 1 – 4.

[23] X. Wei, C.-T. Li, Z. Lei, D. Yi, and S. Z. Li, "Dynamic image-to-class warping for occluded face recognition," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2035 – 2050, Dec. 2014.

[24] X. Wei and C.-T. Li, "Fixation and saccade based face recognition from single image per person with various occlusions and expressions," in *IEEE Conference Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2013, pp. 70 – 75.

[25] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210 – 227, Feb. 2009.

[26] J. Chen, S. Shan, C. He, G. Zhao, M. Pietikainen, X. Chen, and W. Gao, "WLD: A robust local image descriptor," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1705 – 1720, Sep. 2010.

[27] G. Porter and G. Doran, "An anatomical and photographic technique for forensic facial identification," *Forensic science international*, vol. 114, no. 2, pp. 97 – 105, 2000.

[28] A. Bertillon, *La photographie judiciaire, avec un appendice sur la classification et l'identification anthropometriques*. Paris: Gauthier-Villars, 1890.

[29] "The speaking portrait," *Pearson's Magazine*, vol. XI, Jan. - Jun. 1901.

[30] D. Dessimoz and C. Champod, "Linkages between biometrics and forensic science," in *Handbook of Biometrics*, A. Jain, P. Flynn, and A. Ross, Eds. Springer US, 2008, pp. 425 – 459.

[31] T. Ali, R. N. J. Veldhuis, and L. J. Spreeuwers, "Forensic face recognition: A survey," Technical Report TR-CTIT-10-40, Dec. 2010.

[32] P. Sinha, B. Balas, Y. Ostrovsky, and R. Russell, "Face recognition by humans: Nineteen results all computer vision researchers should know about," *Proceedings of the IEEE*, vol. 94, no. 11, pp. 1948 – 1962, 2006.

[33] W.-H. Lai and C.-T. Li, "Skin colour-based face detection in colour images," in *IEEE International Conference Advanced Video and Signal-Based Surveillance (AVSS)*, Nov. 2006, pp. 56 – 61.

[34] A. K. Jain and A. Ross, "Introduction to biometrics," in *Handbook of Biometrics*, A. K. Jain, P. Flynn, and A. A. Ross, Eds. Springer US, 2008, pp. 1 – 22.

[35] B. Arbab-Zavar, X. Wei, J. D. Bustard, M. S. Nixon, and C.-T. Li, "On forensic use of biometrics," in *Handbook of Digital Forensics of Multimedia Data and Devices*, A. T. S. Ho and S. Li, Eds. John Wiley & Sons, Inc., 2014 (in press).

[36] H. Chan and W. Bledsoe, "A man-machine facial recognition system: Some preliminary results," in *Technical Report*, 1965.

[37] T. Kanade, "Picture processing system by computer complex and recognition of human faces," in *Doctoral dissertation, Kyoto University*, Nov. 1973.

[38] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *IEEE Conference Computer Vision and Pattern Recognition (CVPR)*, Jun. 1991, pp. 586 – 591.

[39] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711 – 720, Jul. 1997.

[40] P. S. Penev and J. J. Atick, "Local feature analysis: a general statistical theory for object representation," *Network: Computation in Neural Systems*, vol. 7, no. 3, pp. 477 – 500, 1996.

[41] L. Wiskott, J.-M. Fellous, N. Kuiger, and C. Von der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 775 – 779, 1997.

[42] A. Nefian and I. Hayes, M.H., "Hidden markov models for face recognition," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, vol. 5, May 1998, pp. 2721 – 2724.

[43] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models—their training and application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38 – 59, Jan. 1995.

[44] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," in *European Conference Computer Vision (ECCV)*, ser. Lecture Notes in Computer Science, H. Burkhardt and B. Neumann, Eds. Springer Berlin Heidelberg, 1998, vol. 1407, pp. 484 – 498.

[45] P. J. Grother, G. W. Quinn, and P. J. Phillips, "Report on the evaluation of 2d still-image face recognition algorithms," National Institute of Standards and Technology, Report, Jun. 2010.

[46] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323 – 2326, 2000.

[47] A. Georghiades, P. Belhumeur, and D. Kriegman, "From few to many: illumination cone models for face recognition under variable lighting and pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643 – 660, Jun. 2001.

[48] S. Romdhani, V. Blanz, and T. Vetter, "Face identification by fitting a 3D morphable model using linear shape and texture error functions," in *European Conference Computer Vision (ECCV)*, ser. Lecture Notes in Computer Science, A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, Eds. Springer Berlin Heidelberg, 2002, vol. 2353, pp. 3 – 19.

[49] M. Bartlett, J. R. Movellan, and T. Sejnowski, "Face recognition by independent component analysis," *IEEE Transactions on Neural Networks*, vol. 13, no. 6, pp. 1450 – 1464, Nov. 2002.

[50] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037 – 2041, Dec. 2006.

[51] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation

methodology for face recognition algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090 – 1104, Oct. 2000.

[52] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *IEEE Conference Computer Vision and Pattern Recognition (CVPR)*, vol. 1, Jun. 2005, pp. 947 – 954.

[53] P. Phillips, "The next face challenge: Achieving robust human level performance," in *NIST Biometric Performance Conference*, Mar. 2012.

[54] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar, "Attribute and simile classifiers for face verification," in *IEEE International Conference Computer Vision (ICCV)*, Sep. 2009, pp. 365 – 372.

[55] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137 – 154, 2004.

[56] E. Nowak and F. Jurie, "Learning visual similarity measures for comparing never seen objects," in *IEEE Conference Computer Vision and Pattern Recognition (CVPR)*, Jun. 2007, pp. 1 – 8.

[57] H. Fan, Z. Cao, Y. Jiang, Q. Yin, and C. Doudou, "Learning deep face representation," *CoRR*, vol. abs/1403.2802, 2014.

[58] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *IEEE Conference Computer Vision and Pattern Recognition (CVPR)*, 2014.

[59] C. Lu and X. Tang, "Surpassing human-level face verification performance on LFW with Gaussianface," *CoRR*, vol. abs/1404.3840, 2014.

[60] N. Firth, "Face recognition technology fails to find uk

rioters," http://www.newscientist.com/article/mg21128266.000-face-recognition-technology-fails-to-find-uk-rioters.html/, Aug. 2011.

[61] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Computing Surveys*, vol. 35, no. 4, pp. 399 – 458, Dec. 2003.

[62] A. K. Jain and S. Z. Li, *Handbook of Face Recognition (2nd)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2011.

[63] A. F. Abate, M. Nappi, D. Riccio, and G. Sabatino, "2D and 3D face recognition: A survey," *Pattern Recognition Letters*, vol. 28, no. 14, pp. 1885 – 1906, 2007, image: Information and Control.

[64] X. Zou, J. Kittler, and K. Messer, "Illumination invariant face recognition: A survey," in *IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS)*, Sep. 2007, pp. 1 – 8.

[65] G. Hermosilla, J. R. del Solar, R. Verschae, and M. Correa, "A comparative study of thermal face recognition methods in unconstrained environments," *Pattern Recognition*, vol. 45, no. 7, pp. 2445 – 2459, 2012.

[66] S. Li, R. Chu, S. Liao, and L. Zhang, "Illumination invariant face recognition using near-infrared images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 627 – 639, Apr. 2007.

[67] S. Du and R. Ward, "Wavelet-based illumination normalization for face recognition," in *IEEE International Conference on Image Processing (ICIP)*, vol. 2, Sep. 2005, pp. II–954–7.

[68] S. Shan, W. Gao, B. Cao, and D. Zhao, "Illumination normalization for robust face recognition against varying lighting conditions," in *IEEE International Workshop on Analysis and Modeling of Faces and Gestures (AMFG)*, Oct. 2003, pp. 157 – 164.

[69] T. Chen, W. Yin, X. S. Zhou, D. Comaniciu, and T. Huang, "Total variation models for variable lighting face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 9, pp. 1519 – 1524, Sep. 2006.

[70] Y. Gao and M. K. H. Leung, "Face recognition using line edge map," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 6, pp. 764 – 779, Jun. 2002.

[71] R. Ramamoorthi, "Analytic pca construction for theoretical analysis of lighting variability in images of a lambertian object," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 10, pp. 1322 – 1333, Oct. 2002.

[72] R. Basri and D. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 2, pp. 218 – 233, Feb. 2003.

[73] J. Lambert, "Photometria sive de mensura et gradibus luminus," *Colorum et Umbrae", Eberhard Klett*, 1760.

[74] P. Belhumeur and D. Kriegman, "What is the set of images of an object under all possible lighting conditions?" in *IEEE Conference Computer Vision and Pattern Recognition (CVPR)*, Jun. 1996, pp. 270 – 277.

[75] X. Zhang and Y. Gao, "Face recognition across pose: A review," *Pattern Recognition*, vol. 42, no. 11, pp. 2876 – 2896, 2009.

[76] D. Beymer, "Face recognition under varying pose," in *IEEE Conference Computer Vision and Pattern Recognition (CVPR)*, Jun. 1994, pp. 756 – 761.

[77] R. Singh, M. Vatsa, A. Ross, and A. Noore, "A mosaicing scheme for pose-invariant face recognition," *IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics*, vol. 37, no. 5, pp. 1212 – 1225, Oct. 2007.

[78] D. Beymer and T. Poggio, "Face recognition from one example view," in *IEEE International Conference Computer Vision (ICCV)*, Jun. 1995, pp. 500 – 507.

[79] R. Gross, I. Matthews, and S. Baker, "Appearance-based face recognition and light-fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 4, pp. 449 – 465, Apr. 2004.

[80] D. Gonzalez-Jimenez and J. Alba-Castro, "Toward pose-invariant 2-d face recognition through point distribution models and facial symmetry," *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 3, pp. 413 – 429, Sep. 2007.

[81] T. Kanade and A. Yamada, "Multi-subregion based probabilistic approach toward pose-invariant face recognition," in *IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA)*, vol. 2, Jul. 2003, pp. 954 – 959.

[82] X. Chai, S. Shan, X. Chen, and W. Gao, "Locally linear regression for pose-invariant face recognition," *IEEE Transactions on Image Processing*, vol. 16, no. 7, pp. 1716 – 1725, Jul. 2007.

[83] A. Li, S. Shan, X. Chen, and W. Gao, "Maximizing intra-individual correlations for face recognition across pose differences," in *IEEE Conference Computer Vision and Pattern Recognition (CVPR)*, Jun. 2009, pp. 605 – 611.

[84] V. Blanz and T. Vetter, "Face recognition based on fitting a 3D morphable model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1063 – 1074, Sep. 2003.

[85] X. Liu and T. Chen, "Pose-robust face recognition using geometry assisted probabilistic modeling," in *IEEE Conference Computer Vision and Pattern Recognition (CVPR)*, vol. 1, Jun. 2005, pp. 502 – 509.

[86] X. Zhang, Y. Gao, and M. K. H. Leung, "Automatic texture synthesis for face recognition from single views," in *International Conference on Pattern Recognition (ICPR)*, vol. 3, 2006, pp. 1151 – 1154.

[87] C. Castillo and D. Jacobs, "Using stereo matching with general epipolar geometry for 2d face recognition across pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 12, pp. 2298 – 2304, Dec. 2009.

[88] E. Murphy-Chutorian and M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 607 – 626, Apr. 2009.

[89] A. M. Martínez, "Recognizing expression variant faces from a single sample image per class," in *IEEE Conference Computer Vision and Pattern Recognition (CVPR)*, vol. 1, Jun. 2003, pp. 353– 358.

[90] H. Rashidy Kanan and Y. Gao, "Recognition of expression variant faces from one sample image per enrolled subject," in *IEEE International Conference on Image Processing (ICIP)*, Nov. 2009, pp. 3309 – 3312.

[91] M. Ramachandran, S. Zhou, D. Jhalani, and R. Chellappa, "A method for converting a smiling face to a neutral face with applications to face recognition," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, vol. 2, Mar. 2005, pp. 977 – 980.

[92] C.-K. Hsieh, S.-H. Lai, and Y.-C. Chen, "An optical flow-based approach to robust face recognition under expression variations," *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 233 – 240, 2010.

[93] A. Bronstein, M. Bronstein, and R. Kimmel, "Expression-invariant representations of faces," *IEEE Transactions on Image Processing*, vol. 16, no. 1, pp. 188 – 197, 2007.

[94] D. Smeets, P. Claes, J. Hermans, D. Vandermeulen, and P. Suetens, "A comparative study of 3-D face recognition under expression variations," *IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews*, vol. 42, no. 5, pp. 710 – 727, Sep. 2012.

[95] B. Gunturk, A. Batur, Y. Altunbasak, M. Hayes, and R. Mersereau, "Eigenface-domain super-resolution for face recognition," *IEEE Transactions on Image Processing*, vol. 12, no. 5, pp. 597 – 606, May 2003.

[96] K. Jia and S. Gong, "Multi-modal tensor face for simultaneous super-resolution and recognition," in *IEEE International Conference Computer Vision (ICCV)*, vol. 2, Oct. 2005, pp. 1683 – 1690.

[97] P. Hennings-Yeomans, S. Baker, and B. Kumar, "Simultaneous super-resolution and feature extraction for recognition of low-resolution faces," in *IEEE Conference Computer Vision and Pattern Recognition (CVPR)*, Jun. 2008, pp. 1 – 8.

[98] P. Hennings-Yeomans, B. Kumar, and S. Baker, "Robust low-resolution face identification and verification using high-resolution features," in *IEEE International Conference on Image Processing (ICIP)*, Nov. 2009, pp. 33 – 36.

[99] H. Huang and H. He, "Super-resolution method for face recognition using nonlinear mappings on coherent features," *IEEE Transactions on Neural Networks*, vol. 22, no. 1, pp. 121–130, Jan. 2011.

[100] S.-W. Lee, J. Park, and S.-W. Lee, "Low resolution face recognition based on support vector data description," *Pattern Recognition*, vol. 39, no. 9, pp. 1809 – 1812, 2006.

[101] B. Li, H. Chang, S. Shan, and X. Chen, "Low-resolution face recognition via coupled locality preserving mappings," *IEEE Signal Processing Letters*, vol. 17, no. 1, pp. 20 – 23, Jan. 2010.

[102] S. Biswas, K. Bowyer, and P. Flynn, "Multidimensional scaling for matching low-resolution facial images," in *IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS)*, Sep. 2010, pp. 1 – 6.

[103] S. Shekhar, V. Patel, and R. Chellappa, "Synthesis-based recognition of low resolution faces," in *IEEE International Joint Conference on Biometrics (IJCB)*, Oct. 2011, pp. 1 – 6.

[104] F. Hua, P. Johnson, N. Sazonova, P. Lopez-Meyer, and S. Schuckers, "Impact of out-of-focus blur on face recognition performance based on modular transfer function," in *International Conference Biometrics (ICB)*, Mar. 2012, pp. 85 – 90.

[105] M. Nishiyama, A. Hadid, H. Takeshima, J. Shotton, T. Kozakaya, and O. Yamaguchi, "Facial deblur inference using subspace analysis for recognition of blurred faces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 4, pp. 838 – 845, Apr. 2011.

[106] H. Zhang, J. Yang, Y. Zhang, N. Nasrabadi, and T. Huang, "Close the loop: Joint blind image restoration and recognition with sparse representation prior," in *IEEE International Conference Computer Vision (ICCV)*, Nov. 2011, pp. 770 – 777.

[107] T. Ahonen, E. Rahtu, V. Ojansivu, and J. Heikkila, "Recognition of blurred faces using local phase quantization," in *International Conference on Pattern Recognition (ICPR)*, Dec 2008, pp. 1–4.

[108] A. Hadid, M. Nishiyama, and Y. Sato, "Recognition of blurred faces via facial deblurring combined with blur-tolerant descriptors," in *Pattern Recognition (ICPR), 2010 20th International Conference on*, Aug 2010, pp. 1160–1163.

[109] X. Geng, Z.-H. Zhou, and K. Smith-Miles, "Automatic age estimation based on facial aging patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2234 – 2240, 2007.

[110] G. Guo, Y. Fu, C. Dyer, and T. Huang, "Image-based human age estimation by manifold learning and locally adjusted robust regression," *IEEE Transactions on Image Processing*, vol. 17, no. 7, pp. 1178 – 1188, 2008.

[111] A. Lanitis, C. Taylor, and T. Cootes, "Toward automatic simulation of aging effects on face images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 442 – 455, 2002.

[112] Z. Li, U. Park, and A. Jain, "A discriminative model for age invariant face recognition," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 3, pp. 1028 – 1037, 2011.

[113] M. Yang and L. Zhang, "Gabor feature based sparse representation for face recognition with Gabor occlusion dictionary," in *European Conference Computer Vision (ECCV)*, vol. 6316, 2010, pp. 448 – 461.

[114] Z. Zhou, A. Wagner, H. Mobahi, J. Wright, and Y. Ma, "Face recognition with contiguous occlusion using Markov random fields," in *IEEE International Conference Computer Vision (ICCV)*, Oct. 2009, pp. 1050 – 1057.

[115] C.-F. Chen, C.-P. Wei, and Y.-C. F. Wang, "Low-rank matrix recovery with structural incoherence for robust face recognition," in *IEEE Conference Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 2618 – 2625.

[116] A. I. Naseem, B. R. Togneri, and C. M. Bennamoun, "Linear regression for face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 2106 – 2112, Nov. 2010.

[117] R. He, W.-S. Zheng, and B.-G. Hu, "Maximum correntropy criterion for robust face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1561 – 1576, Aug. 2011.

[118] M. Yang, D. Zhang, J. Yang, and D. Zhang, "Robust sparse coding for face recognition," in *IEEE Conference Computer Vision and Pattern Recognition (CVPR)*, Jun. 2011, pp. 625 – 632.

[119] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in *IEEE International Conference Computer Vision (ICCV)*, 2011, pp. 471 – 478.

[120] X. Tan, S. Chen, Z.-H. Zhou, and F. Zhang, "Recognizing partially occluded, expression variant faces from single training image per person with SOM and soft k-NN ensemble," *IEEE Transactions on Neural Networks*, vol. 16, no. 4, pp. 875 – 886, 2005.

[121] X. Tan, S. Chen, Z.-H. Zhou, and J. Liu, "Face recognition under occlusions and variant expressions with partial similarity," *IEEE Transactions on Information Forensics and Security*, vol. 4, no. 2, pp. 217 – 230, Jun. 2009.

[122] W. Chen and Y. Gao, "Recognizing partially occluded faces from a single sample per class using string-based matching," in *European Conference Computer Vision (ECCV)*, ser. Lecture Notes in Computer Science, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Springer Berlin Heidelberg, 2010, vol. 6313, pp. 496 – 509.

[123] W. Chen and Y. Gao, "Face recognition using ensemble string matching," *IEEE Transactions on Image Processing*, vol. 22, no. 12, pp. 4798 – 4808, Dec. 2013.

[124] R. Weng, J. Lu, J. Hu, G. Yang, and Y.-P. Tan, "Robust feature set matching for partial face recognition," in *IEEE International Conference Computer Vision (ICCV)*, Dec. 2013, pp. 601 – 608.

[125] S. Li, X. Hou, H. Zhang, and Q. Cheng, "Learning spatially localized, parts-based representation," in *IEEE Conference Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2001, pp. 207 – 212.

[126] J. Kim, J. Choi, J. Yi, and M. Turk, "Effective representation using ICA for face recognition robust to local distortion and partial occlusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 12, pp. 1977 – 1981, Dec. 2005.

[127] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang, "Local Gabor binary pattern histogram sequence (LGBPHS): a novel non-statistical model for face representation and recognition," in *IEEE International Conference Computer Vision (ICCV)*, vol. 1, 2005, pp. 786 – 791.

[128] G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "Subspace learning from image gradient orientations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 12, pp. 2454 – 2466, 2012.

[129] J. Zhu, D. Cao, S. Liu, Z. Lei, and S. Z. Li, "Discriminant analysis with Gabor phase for robust face recognition," in *International Conference Biometrics (ICB)*, 2012, pp. 13 – 18.

[130] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91 – 110, 2004.

[131] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Transactions on Image Processing*, vol. 11, no. 4, pp. 467 – 476, Apr. 2002.

[132] A. M. Martínez and R. Benavente, "The AR face database," Tech. Rep., 1998.

[133] K.-C. Lee, J. Ho, and D. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 684 – 698, May 2005.

[134] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild:

A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, Oct. 2007.

[135] D. Miranda, "The face we make," http://www.thefacewemake.org/.

[136] Y. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Transactions on Information Theory*, vol. 55, no. 11, pp. 5302 – 5316, Nov. 2009.

[137] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?" *Vision Research*, vol. 37, no. 23, pp. 3311 – 3325, 1997.

[138] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image Processing*, vol. 15, no. 12, pp. 3736 – 3745, Dec. 2006.

[139] J. Mairal, M. Elad, and G. Sapiro, "Sparse representation for color image restoration," *IEEE Transactions on Image Processing*, vol. 17, no. 1, pp. 53 – 69, Jan. 2008.

[140] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1031 – 1044, Jun. 2010.

[141] E. Amaldi and V. Kann, "On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems," *Theoretical Computer Science*, vol. 209, no. 1-2, pp. 237 – 260, 1998.

[142] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207 – 1223, 2006.

[143] D. L. Donoho, "For most large underdetermined systems of linear equations the

minimal l_1-norm solution is also the sparsest solution," *Communications on Pure and Applied Mathematics*, vol. 59, no. 6, pp. 797 – 829, 2006.

[144] T. Bai and Y. Li, "Robust visual tracking with structured sparse representation appearance model," *Pattern Recognition*, vol. 45, no. 6, pp. 2390 – 2404, 2012, brain Decoding.

[145] X.-T. Yuan and S. Yan, "Visual classification with multi-task joint sparse representation," in *IEEE Conference Computer Vision and Pattern Recognition (CVPR)*, Jun. 2010, pp. 3493 – 3500.

[146] S. Zhang, J. Huang, Y. Huang, Y. Yu, H. Li, and D. Metaxas, "Automatic image annotation using group sparsity," in *IEEE Conference Computer Vision and Pattern Recognition (CVPR)*, Jun. 2010, pp. 3312 – 3319.

[147] Q. Shi, A. Eriksson, A. van den Hengel, and C. Shen, "Is face recognition really a compressive sensing problem?" in *IEEE Conference Computer Vision and Pattern Recognition (CVPR)*, Jun. 2011, pp. 553 – 560.

[148] J. Wright, A. Ganesh, A. Y. Yang, Z. Zhou, and Y. Ma, "Sparsity and robustness in face recognition," *CoRR*, vol. abs/1111.1014, 2011.

[149] C. Ding, D. Zhou, X. He, and H. Zha, "R1-PCA: rotational invariant L1-norm principal component analysis for robust subspace factorization," in *International Conference on Machine learning (ICML)*, 2006, pp. 281 – 288.

[150] E. Elhamifar and R. Vidal, "Robust classification using structured sparse representation," in *IEEE Conference Computer Vision and Pattern Recognition (CVPR)*, Jun. 2011, pp. 1873 – 1879.

[151] K. Jia, T.-H. Chan, and Y. Ma, "Robust and practical face recognition via structured sparsity," in *European Conference Computer Vision (ECCV)*, ser. Lecture Notes in

Computer Science, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Springer Berlin Heidelberg, 2012, vol. 7575, pp. 331 – 344.

[152] Y. Eldar and H. Bolcskei, "Block-sparsity: Coherence and efficient recovery," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Apr. 2009, pp. 2885 – 2888.

[153] A. Majumdar and R. Ward, "Non-convex group sparsity: Application to color imaging," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Mar. 2010, pp. 469 – 472.

[154] A. Jain, "Fundamentals of digital signal processing," *Fundamentals of Digital Signal Processing*, 1989.

[155] Z. Hou and W.-Y. Yau, "Relative gradients for image lighting correction," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Mar. 2010, pp. 1374 – 1377.

[156] T. Zhang, Y.-Y. Tang, B. Fang, Z. Shang, and X. Liu, "Face recognition under varying illumination using gradientfaces," *IEEE Transactions on Image Processing*, vol. 18, no. 11, pp. 2599 – 2606, Nov. 2009.

[157] B. Wang, W. Li, W. Yang, and Q. Liao, "Illumination normalization based on Weber's law with application to face recognition," *IEEE Signal Processing Letters*, vol. 18, no. 8, pp. 462 – 465, Aug. 2011.

[158] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using laplacianfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 328 – 340, Mar. 2005.

[159] A. M. Martínez and A. Kak, "PCA versus LDA," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 228 – 233, Feb. 2001.

[160] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43 – 49, Feb. 1978.

[161] S. U. Hussain, T. Napoléon, and F. Jurie, "Face recognition using local quantized patterns," in *British Machine Vision Conference (BMVC)*, 2012, pp. 99.1 – 99.11.

[162] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1 – 27:27, 2011, http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[163] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification," in *IEEE Conference Computer Vision and Pattern Recognition (CVPR)*, Jun. 2008, pp. 1 – 8.

[164] H. Bay, T. Tuytelaars, and L. Gool, "SURF: Speeded up robust features," in *European Conference Computer Vision (ECCV)*, vol. 3951, 2006, pp. 404 – 417.

[165] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE Conference Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2006, pp. 2169 – 2178.

[166] L. Pishchulin, T. Gass, P. Dreuw, and H. Ney, "Image warping for face recognition: From local optimality towards global optimization," *Pattern Recognition*, vol. 45, no. 9, pp. 3131 – 3140, 2012.

[167] K. Tan and S. Chen, "Adaptively weighted sub-pattern PCA for face recognition," *Neurocomputing*, vol. 64, pp. 505 – 511, 2005.

[168] P. Dreuw, P. Steingrube, H. Hanselmann, and H. Ney, "SURF-face: Face recognition under viewpoint consistency constraints," in *British Machine Vision Conference (BMVC)*, 2009, pp. 7.1 – 7.11.

[169] H. J. Seo and P. Milanfar, "Face verification using the LARK representation," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 4, pp. 1275 – 1286, 2011.

[170] G. Sharma, S. Hussain, and F. Jurie, "Local higher-order statistics (LHS) for texture categorization and facial analysis," in *European Conference Computer Vision (ECCV)*, vol. 7578, 2012, pp. 1 – 12.

[171] O. Barkan, J. Weill, L. Wolf, and H. Aronowitz, "Fast high dimensional vector multiplication face recognition," in *IEEE International Conference Computer Vision (ICCV)*, Dec. 2013, pp. 1960 – 1967.

[172] J. Ruiz-del Solar, R. Verschae, and M. Correa, "Recognition of faces in unconstrained environments: A comparative study," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, pp. 1:1 – 1:19, Jan. 2009.

[173] D. Chen, X. Cao, F. Wen, and J. Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *IEEE Conference Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 3025 – 3032.

[174] S. Z. Li, D. Yi, Z. Lei, and S. Liao, "The CASIA NIR-VIS 2.0 face database," in *IEEE Conference Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2013, pp. 348 – 353.

[175] R. Wang, Z. Lei, M. Ao, and S. Z. Li, "Bayesian face recognition based on markov random field modeling," in *International Conference Biometrics (ICB)*, vol. 5558, 2009, pp. 42 – 51.

[176] D. Keysers and W. Unger, "Elastic image matching is NP-complete," *Pattern Recognition Letters*, vol. 24, pp. 445 – 453, 2003.

[177] J. Lu, Y.-P. Tan, and G. Wang, "Discriminative multimanifold analysis for face

recognition from a single training sample per person," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 39 – 51, 2013.

[178] L. Barrington, T. K. Marks, J. Hui-wen Hsiao, and G. W. Cottrell, "NIMBLE: A kernel density model of saccade-based visual memory," *Journal of Vision*, vol. 8, no. 14, 2008.

[179] C. Kanan and G. Cottrell, "Robust classification of objects, faces, and flowers using natural image statistics," in *IEEE Conference Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 2472 – 2479.

[180] M. De Marsico, M. Nappi, and D. Riccio, "FARO: Face recognition against occlusions and expression variations," *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, vol. 40, no. 1, pp. 121 – 132, 2010.

[181] T. Gass, L. Pishchulin, P. Dreuw, and H. Ney, "Warp that smile on your face: Optimal and smooth deformations for face recognition," in *IEEE International Conference Automatic Face and Gesture Recognition (FG)*, 2011, pp. 456 – 463.

[182] Y. Guan, X. Wei, C.-T. Li, G. Marcialis, F. Roli, and M. Tistarelli, "Combining gait and face for tackling the elapsed time challenges," in *IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS)*, Sep. 2013, pp. 1 – 8.

[183] P. Tresadern, T. Cootes, N. Poh, P. Matejka, A. Hadid, C. Levy, C. McCool, and S. Marcel, "Mobile biometrics: Combined face and voice verification for a mobile platform," *IEEE Pervasive Computing*, vol. 12, no. 1, pp. 79–87, Jan. 2013.